



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0069502
(43) 공개일자 2021년06월11일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06F 17/15 (2006.01)
G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/082 (2013.01)
G06F 17/153 (2013.01)
(21) 출원번호 10-2019-0159407
(22) 출원일자 2019년12월03일
심사청구일자 2019년12월03일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
배성호
경기도 용인시 기흥구 예현로 15 서그내마을에스케이아파트 102동 501호
정준현
서울특별시 서초구 남부순환로350길 29-10, B01호 (양재동)
(74) 대리인
특허법인지원

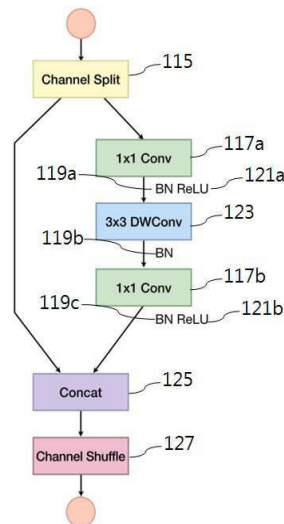
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 인공 신경망 네트워크에서 고속 주파수 변환을 이용한 점별 컨벌루션을 위한 장치 및 그 제어 방법

(57) 요약

다양한 실시예들에 따르면, 컨벌루션 신경망 네트워크는, 입력 데이터를 수신하도록 설정된 입력 계층, 수신된 입력 데이터에 대응하는 출력 데이터를 출력하도록 설정된 출력 계층 및 입력 계층 및 출력 계층 사이에 위치하는 복수의 은닉 계층들을 포함하고, 복수의 은닉 계층들 중 적어도 하나는, 이전의 계층적 단계에서 추출된 이전의 특성 정보에 기반하여 채널 방향에서의 고속화 변환 연산을 수행하여 특성 정보를 추출하도록 설정된 컨벌루션 네트워크 모듈을 포함할 수 있다.

대표도 - 도2a



(52) CPC특허분류

G06N 3/04 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	10085646
부처명	과학기술정보통신부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	전자정보디바이스산업원천기술개발사업(반도체)
연구과제명	3D 맵리스터 어레이 결점 극복을 위한 뉴로모픽 시스템
기 여 율	1/2
과제수행기관명	경희대학교산학협력단
연구기간	2019.06.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	2015-0-00742
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	Grand ICT연구센터 지원사업
연구과제명	라이프 컴패니온쉽 경험을 위한 지능형 인터렉션 융합 연구
기 여 율	1/2
과제수행기관명	성균관대학교 산학협력단
연구기간	2019.01.01 ~ 2019.12.31

명세서

청구범위

청구항 1

컨벌루션 신경망 네트워크에 있어서,

입력 데이터를 수신하도록 설정된 입력 계층;

상기 수신된 입력 데이터에 대응하는 출력 데이터를 출력하도록 설정된 출력 계층; 및

상기 입력 계층 및 상기 출력 계층 사이에 위치하는 복수의 은닉 계층들을 포함하고,

상기 복수의 은닉 계층들 중 적어도 하나는,

이전의 계층적 단계에서 추출된 이전의 특성 정보에 기반하여 채널 방향에서의 고속화 변환 연산을 수행하여 특성 정보를 추출하도록 설정된 컨벌루션 네트워크 모듈을 포함하는 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 2

제1항에 있어서,

상기 컨벌루션 네트워크 모듈은,

학습이 필요 없는 가중치들을 이용하여 상기 채널 방향에서의 고속화 변환 연산을 수행하도록 설정된 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 3

제1항에 있어서,

상기 복수의 은닉 계층들은,

상기 입력 데이터에 기반하여 낮은 계층적 단계에서의 제1 특성 정보를 추출하도록 설정된 제1 컨벌루션 네트워크 모듈,

상기 제1 특성 정보에 기반하여 중간 계층적 단계에서의 제2 특성 정보를 추출하도록 설정된 제2 컨벌루션 네트워크 모듈, 및

상기 제2 특성 정보에 기반하여 높은 계층적 단계에서의 제3 특성 정보를 추출하도록 설정된 제3 컨벌루션 네트워크 모듈을 포함하고,

상기 제3 컨벌루션 네트워크 모듈은, 상기 제2 특성 정보에 대하여 상기 채널 방향에서의 고속화 변환 연산을 수행하여 상기 제3 특성 정보를 추출하도록 설정된 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 4

제3항에 있어서,

상기 제1 컨벌루션 네트워크 모듈은, 상기 입력 데이터에 대하여 공간 방향에서의 분리 가능한 컨벌루션 연산 및 채널 방향에서의 일차원 점별 컨벌루션 연산을 수행하여 상기 제1 특성 정보를 추출하도록 설정되고,

상기 제2 컨벌루션 네트워크 모듈은, 상기 제1 특성 정보에 대하여 공간 방향에서의 분리 가능한 컨벌루션 연산 및 채널 방향에서의 일차원 점별 컨벌루션을 수행하여 상기 제2 특성 정보를 추출하도록 설정된 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 5

제2항에 있어서,

상기 제3 컨벌루션 네트워크 모듈은, 상기 제2 특성 정보에 대하여 공간 방향에서의 분리 가능한 컨벌루션 연산을 수행하도록 더 설정된 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 6

제3항에 있어서,

상기 제3 컨벌루션 네트워크 모듈은, 채널 분리 연산, 배치 단위의 정규화 연산, 채널 통합 연산 또는 채널 혼합 연산 중 적어도 하나를 수행하도록 더 설정된 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 7

제1항에 있어서,

상기 채널 방향에서의 고속화 변환 연산은,

일차원 이산 윌시-하다마드 변환 연산 또는 일차원 이산 코사인 변환 연산 중 적어도 하나를 포함하는 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 8

제7항에 있어서,

상기 일차원 이산 윌시-하다마드 변환 연산은,

$Z_{ijm} = H_m^D \cdot X_{ij}$, $1 \leq m \leq M$ 에 따라 정의되는 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 9

제7항에 있어서,

상기 일차원 이산 코사인 변환 연산은,

$Z_{ijm} = C_m^D \cdot X_{ij}$, $1 \leq m \leq M$ 에 따라 정의되는 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 10

제7항에 있어서,

상기 일차원 이산 윌시-하다마드 변환 연산은,

고속화 DWT 알고리즘에 기반한 연산을 포함하는 것을 특징으로 하는 컨벌루션 신경망 네트워크.

청구항 11

전자 장치에 있어서,

메모리; 및

적어도 하나의 프로세서를 포함하고,

상기 메모리는, 실행될 때, 상기 적어도 하나의 프로세서가:

입력 특징 맵을 수신하고

상기 입력 특징 맵의 개수가 출력 특징 맵의 개수보다 큰지 여부를 판단하고,

상기 입력 특징 맵의 개수가 상기 출력 특징 맵의 개수보다 크다고 판단되면, 짝수번째 인덱스를 가지는 채널 요소들과 홀수번째 인덱스를 가지는 채널 요소들에 대하여, 요소별로 덧셈 연산을 한 값과 뺄셈 연산을 한 값은 각각 앞에서부터 N/2번째의 채널 요소들과 N/2번째부터 마지막 번째까지의 채널 요소들로 대체하도록 하는 인스트럭션들을 저장하도록 설정된 것을 특징으로 하는 전자 장치.

청구항 12

제11항에 있어서,

상기 인스트럭션들은, 상기 적어도 하나의 프로세서가:

상기 입력 특징 맵의 개수가 상기 출력 특징 맵의 개수보다 크다고 판단되면, 채널 축에 대하여 상기 입력 특징 맵에 대하여 0값을 패딩하도록 하는 것을 특징으로 하는 전자 장치.

청구항 13

제11항에 있어서,

상기 인스트럭션들은, 상기 적어도 하나의 프로세서가:

상기 입력 특징 맵의 개수보다 상기 출력 특징 맵의 개수가 작다고 판단되면, 채널 요소에 대하여 앞에서부터 출력 특징 맵 개수까지를 제외한 채널 요소들을 절삭하도록 하는 것을 특징으로 하는 전자 장치.

청구항 14

전자 장치를 제어하는 방법에 있어서,

입력 특징 맵을 수신하는 단계;와

상기 입력 특징 맵의 개수가 출력 특징 맵 개수보다 큰지 여부를 판단하는 단계와;

상기 입력 특징 맵의 개수가 출력 특징 맵 개수보다 크다고 판단되면, 짝수번째 인덱스를 가지는 채널 요소들과 홀수번째 인덱스를 가지는 채널 요소들에 대하여, 요소별로 덧셈 연산을 한 값과 뺄셈 연산을 한 값은 각각 앞에서부터 N/2번째의 채널 요소들과 N/2번째부터 마지막 번째까지의 채널 요소들로 대체하는 단계를 포함하는 것을 특징으로 하는 방법.

청구항 15

제14항에 있어서,

상기 입력 특징 맵의 개수가 상기 출력 특징 맵 개수보다 크다고 판단되면, 채널 축에 대하여 상기 입력 특징 맵에 대하여 0값을 패딩하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 16

제14항에 있어서,

입력 특징 맵의 개수보다 출력 특징 맵의 개수가 작다고 판단되면, 채널 요소에 대하여 앞에서부터 출력 특징 맵 개수까지를 제외한 채널 요소들을 절삭하는 단계를 더 포함하는 것을 특징으로 하는 방법.

발명의 설명

기술 분야

[0001] 본 개시의 다양한 실시예들은, 인공 신경망 네트워크에서 고속 주파수 변환을 이용한 점별 컨벌루션을 위한 장치 및 그 제어 방법에 관한 것이다.

배경 기술

[0002] 컨벌루션 신경망 네트워크(convolutional neural network)는, AlexNet이 ImageNet Challenge에서 우승한 이래로 인공 지능 분야(예: 이미지 분류 분야)에서 널리 사용되고 있다. 컨벌루션 신경망 네트워크에 있어서, 이미지 분류 정확도를 높이기 위하여 레이어(layer)들의 깊이를 증가시키는 것이 주된 관심사였다. 하지만, 이미지 분류 정확도를 높이기 위하여 레이어들의 깊이를 증가시킬 경우, 정확도는 높아지는 반면, 연산 비용(예: 연산량 또는 연산 속도) 측면에서 큰 손실이 발생하는 트레이드-오프(trade-off)가 발생할 수 있다.

[0003] 상술한 트레이드-오프 관점에서 효율을 증가시키기 위하여, 다양한 연구가 선행되었다. 예를 들어, Howard et al(Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017.)은, 각 커널에 대하여 3 x 3 x C 크기를 가지는 컨벌루

선을 $3 \times 3 \times 1$ 크기를 가지는 공간 방향에서의 컨벌루션(spatial information specific depth-wise convolution)과 $1 \times 1 \times C$ 크기를 가지는 채널 방향의 점별 컨벌루션(channel information specific point-wise convolution)으로 분리하여 적용하는 방안을 제안함으로써, 이미지 분류 정확도를 유지하면서도 가중치 값 개수와 연산 비용을 줄이고자 하였다.

발명의 내용

해결하려는 과제

[0004] 상술한 채널 방향의 점별 컨벌루션은, 전체 네트워크에서의 연산량(예: FLOPs(floating-point operations))과 파라미터(parameter) 측면에서 높은 비중(예: 각각 약 94.86% 및 74%)을 차지하고 있다. 이에 따른 높은 계산 복잡도와 많은 메모리 공간에 대한 요구로 인하여, 모바일 기기 등 연산 속도와 메모리 공간에 제약이 있는 전자 장치에 대하여 제한적인 적용만 가능하다는 한계가 있을 수 있다.

[0005] 다양한 실시예들에 따라서, 파라미터를 요구하지 않고 고속화 신호 처리 기법에 기반한 새로운 점별 컨벌루션 모듈(다른 말로, 점별 컨벌루션 레이어(PC layer))이 제공될 수 있다.

과제의 해결 수단

[0006] 다양한 실시예들에 따르면, 컨벌루션 신경망 네트워크는, 입력 데이터를 수신하도록 설정된 입력 계층, 수신된 입력 데이터에 대응하는 출력 데이터를 출력하도록 설정된 출력 계층 및 입력 계층 및 출력 계층 사이에 위치하는 복수의 은닉 계층들을 포함하고, 복수의 은닉 계층들은, 컨벌루션을 수행할 수 있다. 이때, 컨벌루션 연산에 점별 컨벌루션을 수행할 경우, 제안 방법은 수행되는 점별 컨벌루션의 가중치를 학습하지 않고 기존의 주파수 변환 커널(예를 들어, 월시-하다마드 변환, 이산코사인 변환, 이산 웨이블릿 변환 등)을 이용해서 수행할 수 있다. 이때, 고속 주파수 변환 기법(예를 들어, 고속 월시-하다마드 변환, 고속 이산코사인 변환, 고속 이산 웨이블릿 변환 등)으로 점별 컨벌루션을 대체할 수 있다.

[0007] 다양한 실시예들에 따르면, 컨벌루션 신경망 네트워크는, 입력 데이터를 수신하도록 설정된 입력 계층, 수신된 입력 데이터에 대응하는 출력 데이터를 출력하도록 설정된 출력 계층 및 입력 계층 및 출력 계층 사이에 위치하는 복수의 은닉 계층들을 포함하고, 복수의 은닉 계층들 중 적어도 하나는, 이전의 계층적 단계에서 추출된 이전의 특성 정보에 기반하여 채널 방향에서의 고속화 변환 연산을 수행하여 특성 정보를 추출하도록 설정된 컨벌루션 네트워크 모듈을 포함할 수 있다.

[0008] 다양한 실시예들에 따르면, 전자 장치는, 메모리 및 적어도 하나의 프로세서를 포함하고, 메모리는, 실행될 때, 적어도 하나의 프로세서가 입력 특징 맵을 수신하고, 입력 특징 맵의 개수가 출력 특징 맵의 개수보다 큰지 여부를 판단하고, 입력 특징 맵의 개수가 출력 특징 맵의 개수보다 크다고 판단되면, 짝수번째 인덱스를 가지는 채널 요소들과 홀수번째 인덱스를 가지는 채널 요소들에 대하여, 요소별로 덧셈 연산을 한 값과 뺄셈 연산을 한 값은 각각 앞에서부터 N/2번째의 채널 요소들과 N/2번째부터 마지막 번째까지의 채널 요소들로 대체하도록 하는 인스트럭션들을 저장하도록 설정될 수 있다.

[0009] 다양한 실시예들에 따르면, 전자 장치를 제어하는 방법은, 입력 특징 맵을 수신하는 단계와 입력 특징 맵의 개수가 출력 특징 맵 개수보다 큰지 여부를 판단하는 단계와 입력 특징 맵의 개수가 출력 특징 맵 개수보다 크다고 판단되면, 짝수번째 인덱스를 가지는 채널 요소들과 홀수번째 인덱스를 가지는 채널 요소들에 대하여, 요소별로 덧셈 연산을 한 값과 뺄셈 연산을 한 값은 각각 앞에서부터 N/2번째의 채널 요소들과 N/2번째부터 마지막 번째까지의 채널 요소들로 대체하는 단계를 포함할 수 있다.

발명의 효과

[0010] 다양한 실시예들에 따른 전자 장치는, 새로운 점별 컨벌루션 모듈을 이용하여, 전체 신경망 네트워크에서의 연산량을 감소시킬 수 있다.

[0011] 다양한 실시예들에 따른 전자 장치는, 파라미터를 요구하지 않고 고속화가 가능한 신호 처리 기법들의 기저들(bases)을 이용하여, 정확도를 유지하면서도 전체 신경망 네트워크에서의 연산 복잡도를 감소시킬 수 있다.

도면의 간단한 설명

[0012] 도 1a은, 종래의 공간 방향의 분리 가능한 컨벌루션을 설명하기 위한 예시 도면이다.

도 1b는, 종래의 공간 방향의 분리 가능한 컨벌루션에 관한 블록 구조(block structure)를 설명하기 위한 예시 도면이다.

도 2a는, 다양한 실시예들에 따른, 고속화 신호 처리 기법에 기반한 점별 컨벌루션 모듈을 포함하는 블록 구조를 설명하기 위한 예시 도면이다.

도 2b는, 다양한 실시예들에 따른, 종래의 채널 방향의 점별 컨벌루션 모듈, 이산 코사인 변환 기반의 CTPC 모듈 및 이산 윌시-하다마드 변환 기반의 CTPC 모듈의 곱셈 연산의 수를 비교하기 위한 그래프이다.

도 3은, 다양한 실시예들에 따른, 이산 윌시-하다마드 변환을 이용한 CTPC 모듈의 동작 방법을 설명하기 위한 알고리즘이다.

도 4a는, 다양한 실시예들에 따른, 점별 컨벌루션 모듈을 포함하는 블록 구조들을 설명하기 위한 예시 도면이다.

도 4b는, 다양한 실시예들에 따른, 본원 발명의 점별 컨벌루션 모듈이 적용되는 신경망 네트워크 구조에서의 계층 레벨(hierarchy level)에 따른 성능을 비교하기 위한 그래프들이다.

도 5는, 본원 발명에 대한 ReLU 활성화 함수의 영향을 설명하기 위한 그래프이다.

도 6은, 본원 발명의 높은 정확도를 설명하기 위한 그래프이다.

도 7은, 높은 계층 레벨에서 3 x 3 공간 방향의 컨벌루션 가중치들의 영향을 설명하기 위한 그래프이다.

발명을 실시하기 위한 구체적인 내용

도 1a는, 종래의 공간 방향의 분리 가능한 컨벌루션을 설명하기 위한 예시 도면이다.

도 1a의 (a)는, 종래의 컨벌루션 필터(다른 말로, 커널(kernel) 또는 레이어(layer))를 도시한다.

종래의 컨벌루션 필터(101)는, 입력(F)으로 크기가 $D_F \times D_F \times M$ 인 특성 맵(feature map)(다른 말로, 특성 벡터 또는 특성 정보)을 수신하고, 출력(G)으로 크기가 $D_F \times D_F \times N$ 인 특성 맵을 출력할 수 있다. 종래의 컨벌루션 필터(101)는 크기가 $D_K \times D_K \times M$ 인 N 개의 필터(103)로 구성될 수 있다. 여기서, D_F 는 입력의 공간 폭(spatial width) 및 공간 높이(spatial height)를 의미할 수 있다. M 은 입력 채널의 수(다른 말로, 입력 깊이(input depth))를 의미할 수 있다. N 은 출력 채널의 수(다른 말로, 출력 깊이(output depth))를 의미할 수 있다.

종래의 컨벌루션 필터(101)를 이용한 연산 비용은 아래와 같이 계산될 수 있다.

수학식 1

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$$

도 1a의 (b)는, 종래의 공간 방향의 컨벌루션 필터(105)를 설명하기 위한 예시 도면이다. 도 1a의 (c)는, 종래의 채널 방향의 점별 컨벌루션 필터(109)를 설명하기 위한 예시 도면이다. 공간 방향의 분리 가능한 컨벌루션 필터는, 공간 방향의 컨벌루션 필터(105) 및 채널 방향의 점별 컨벌루션 필터(109)로 구성될 수 있다.

종래의 공간 방향의 컨벌루션 필터(105)는, 채널당 하나의 필터 적용하는 것으로, 크기가 $D_K \times D_K \times 1$ 인 M 개의 필터(107)로 정의될 수 있다.

종래의 공간 방향의 컨벌루션 필터(105)는 오직 입력 채널만을 출력하기 때문에, 채널 방향의 점별 컨벌루션 필터(109)를 이용하여, 공간 방향의 컨벌루션 필터(105)의 출력들을 선형 조합(linear combination)하여 출력할 수 있다.

종래의 채널 방향의 점별 컨벌루션 필터(109)는, 크기가 $1 \times 1 \times M$ 인 N 개의 필터(111)로 정의될 수 있다.

종래의 공간 방향의 분리 가능한 컨벌루션 필터를 이용한 연산 비용은 아래와 같이 계산될 수 있다.

수학식 2

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$$

[0023]

[0024]

여기서, $D_K \times D_K \times M \times D_F \times D_F$ 는 공간 방향의 컨벌루션 필터(105)의 연산 비용이고, $M \times N \times D_F \times D_F$ 는 채널 방향의 점별 컨벌루션 필터(109)의 연산 비용이다.

[0025]

수학식 1 및 수학식 2를 비교할 때, 공간 방향의 분리 가능한 컨벌루션 필터의 연산 비용(수학식 2)이 종래의 컨벌루션 필터(101)의 연산 비용(수학식 1)보다 작다는 것을 알 수 있다.

[0026]

하지만, 종래의 채널 방향의 점별 컨벌루션 필터(105)의 연산량은, 공간 방향의 분리 가능한 컨벌루션 필터의 연산량의 대부분(예: 약 94%)을 차지한다.

[0027]

따라서 본 개시에서는, 종래의 공간 방향의 컨벌루션 연산의 효율을 증가시키기 위하여 채널 방향의 점별 컨벌루션에 따른 연산량을 낮출 수 있는 방안을 제안하고자 한다.

[0028]

도 1b는, 종래의 공간 방향의 분리 가능한 컨벌루션에 관한 블록 구조(block structure)를 설명하기 위한 예시 도면이다.

[0029]

종래의 공간 방향의 분리 가능한 컨벌루션에 관한 블록 구조는, 채널 분리 모듈(115), 채널 방향의 점별 컨벌루션 모듈(117a, 117b), 배치 단위의 정규화 모듈(119a, 119b, 119c), ReLU(121a, 121b), 공간 방향의 컨벌루션 모듈(123), 채널 통합 모듈(125) 또는 채널 혼합 모듈(127) 중 적어도 하나를 포함할 수 있다.

[0030]

상술한 모듈들 중 적어도 하나는, Ma et al.(Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet V2: practical guidelines for efficient CNN architecture design. CoRR, abs/1807.11164, 2018.)에서 제안하는 모듈을 포함할 수 있다.

[0031]

채널 방향의 점별 컨벌루션 모듈(117a, 117b)은, 채널 방향의 점별 컨벌루션 필터(예: 도 1a의 채널 방향의 점별 컨벌루션 필터(105))를 통한 연산을 수행하도록 설정될 수 있다.

[0032]

공간 방향의 컨벌루션 모듈(123)은, 공간 방향의 컨벌루션 필터(예: 도 1a의 공간 방향의 컨벌루션 필터(111))를 통한 연산을 수행하도록 설정될 수 있다.

[0033]

도 2a는, 다양한 실시예들에 따른, 고속화 신호 처리 기법에 기반한 점별 컨벌루션 모듈을 포함하는 블록 구조를 설명하기 위한 예시 도면이다.

[0034]

다양한 실시예들에 따른 본원 발명의 블록 구조는, 도 1b에 도시된 종래의 공간 방향의 분리 가능한 컨벌루션에 관한 블록 구조와 비교할 때, 채널 방향의 점별 컨벌루션 모듈(117a, 117b)이 CTPC(conventional transform pointwise convolution) 모듈(201a, 201b)로 대체되고, ReLU(121a, 121b)이 생략된 구조일 수 있다. 이하에서는, 도 1b와 중복되는 설명은 생략하여 본원 발명의 블록 구조를 설명하도록 한다.

[0035]

본 개시에서 언급된 “모듈”은 전자 장치의 프로세서의 일부를 구성하거나, 또는, “모듈”에 의해 수행되는 동작들은 프로세서에 의해 수행되는 동작을 의미할 수 있다.

[0036]

본 개시에서 언급된 “모듈” 또는 “모듈”에 의해 수행되는 동작들은, 전자 장치의 메모리에 저장된 명령어(예: 인스트럭션)로 구현될 수 있다.

[0037]

프로세서는, 예를 들면, 소프트웨어(예: 프로그램)를 실행하여 프로세서에 연결된 전자 장치의 적어도 하나의 다른 구성요소(예: 하드웨어 또는 소프트웨어 구성요소)를 제어할 수 있고, 다양한 데이터 처리 또는 연산을 수행할 수 있다. 일 실시예에 따르면, 데이터 처리 또는 연산의 적어도 일부로서, 프로세서(120)는 다른 구성요소(예: 센서 모듈 또는 통신 모듈)로부터 수신된 명령 또는 데이터를 휘발성 메모리에 로드하고, 휘발성 메모리에 저장된 명령 또는 데이터를 처리하고, 결과 데이터를 비휘발성 메모리에 저장할 수 있다. 일 실시예에 따르면, 프로세서는 메인 프로세서(예: 중앙 처리 장치 또는 어플리케이션 프로세서), 및 이와는 독립적으로 또는 함께 운영 가능한 보조 프로세서(예: 그래픽 처리 장치, 이미지 시그널 프로세서, 센서 허브 프로세서, 또는 커뮤니케이션 프로세서)를 포함할 수 있다. 추가적으로 또는 대체적으로, 보조 프로세서는 메인 프로세서보다 저전력을 사용하거나, 또는 지정된 기능에 특화되도록 설정될 수 있다. 보조 프로세서는 메인 프로세서와 별개로, 또

는 그 일부로서 구현될 수 있다.

[0038] 메모리는, 전자 장치의 적어도 하나의 구성요소(예: 프로세서 또는 센서모듈)에 의해 사용되는 다양한 데이터를 저장할 수 있다. 데이터는, 예를 들어, 소프트웨어(예: 프로그램) 및, 이와 관련된 명령에 대한 입력 데이터 또는 출력 데이터를 포함할 수 있다. 메모리는, 휘발성 메모리 또는 비휘발성 메모리를 포함할 수 있다.

[0039] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, 이산 코사인 변환(discrete cosine transform(DCT)) 또는 이산 월시-하다마드 변환(discrete walsh hadamard transform(DWHT)) 중 적어도 하나를 수행하도록 설정될 수 있다.

[0040] 이산 코사인 변환은, N-point input sequence에 대하여, 수학식 3과 같은 기저 커널(basis kernel)을 이용할 수 있다.

수학식 3

$$C_m = [\cos(\frac{(2x+1) \cdot m \cdot \pi}{2N})], \quad 0 \leq x \leq N-1$$

[0042] 여기서, m은 기저의 인덱스를 의미하고, m이 증가함에 따라 입력 신호에서 더 높은 주파수 정보가 포착(capture)되는 특성을 가질 수 있다. 이러한 특성으로 인하여, 이산 코사인 변환은, 낮은 주파수 대역에서의 이미지 신호의 파워를 강조하는 이미지/비디오 압축 기술에서 널리 적용될 수 있다.

[0043] 이산 월시-하다마드 변환은, 커널들에서 +1 및 -1의 값만을 사용하는 빠르고 효율적인 변환이다. 이산 월시-하다마드 변환은, 커널들에서 이진 요소(binary element)만을 포함하므로, 곱셈 연산 없이 합셈 또는 뺄셈 연산만을 수행할 수 있다. 따라서, 이산 월시-하다마드 변환은 텍스처 영상 분할(texture image segmentation), 얼굴 인식, 비디오 샷 경계 검출(video shot boundary detection)과 같은 분야에서 빠른 특성(feature) 추출을 위해 사용될 수 있다.

[0044] 이산 월시-하다마드 변환은, 수학식 4와 같은 기저 커널 행렬을 이용할 수 있다.

수학식 4

$$H^D = \begin{pmatrix} H^{D-1} & H^{D-1} \\ H^{D-1} & -H^{D-1} \end{pmatrix}$$

[0046] 여기서, H₀=1이고, D는 1이상의 값일 수 있다. 본 개시에서의 H_m^D은 H^D의 m번째 행 벡터(row vector)를 의미할 수 있다.

[0047] 종래의 채널 방향의 점별 컨벌루션 모듈(117a, 117b)은, 수학식 5와 같은 연산을 통하여, 입력 X_{ij}에 대한 출력 Z_{ijm}을 출력할 수 있다.

수학식 5

$$Z_{ijm} = W_m^T \cdot X_{ij}, \quad 1 \leq m \leq M$$

[0049] 여기서, (i, j)는 공간 인덱스를 의미하고, m은 출력 채널 인덱스를 의미할 수 있다. N은 입력 채널들의 수를 의미하고, M은 출력 채널들의 수를 의미할 수 있다. X_{ij}는 입력 X의 공간 인덱스 (i, j)에서의 벡터이고, W_m은 가중치 W의 m번째 벡터일 수 있다. 상술한 수학식 5에서는, 설명의 편의상 스트라이드(stride)를 1로 설정하였다.

[0050] 다양한 실시예들에 따라서, CTPC 모듈(201a, 201b)은, 수학식 5의 학습 가능한 파라미터 W_m 대신에, 수학식 3의

C_m 또는 수학적 식 4의 H_m^D 으로 대체한 수학적식에 따른 연산을 수행할 수 있다.

- [0051] 도 2b는, 다양한 실시예들에 따른, 종래의 채널 방향의 점별 컨벌루션 모듈, 이산 코사인 변환 기반의 CTPC 모듈 및 이산 윌시-하다마드 변환 기반의 CTPC 모듈의 곱셈 연산의 수를 비교하기 위한 그래프이다.
- [0052] 도 2b의 그래프를 참조할 때, x축(가로 축)은 입력 채널의 수($2n$)를 의미하고, y축(세로 축)은 곱셈 연산(예: 부동 소수점 곱셈 연산)의 수를 의미한다.
- [0053] 선 203, 205, 207은 각각, 입력 채널의 수($2n$)를 기준으로, Conventional PC(예: 도 2a의 종래의 채널 방향의 점별 컨벌루션 모듈)의 곱셈 연산의 수, PC based on fast DWHT(예: 도 2a의 이산 윌시-하다마드 변환 기반의 CTPC 모듈)의 곱셈 연산의 수, PC based on fast DCT(예: 이산 코사인 변환 기반의 CTPC 모듈)의 곱셈 연산의 수를 의미한다.
- [0054] 선 203과, 선 205 및 207을 비교할 때, Conventional PC(예: 도 2a의 종래의 채널 방향의 점별 컨벌루션 모듈)의 곱셈 연산의 수에 비해, PC based on fast DWHT(예: 도 2a의 이산 윌시-하다마드 변환 기반의 CTPC 모듈)의 곱셈 연산의 수와 PC based on fast DCT(예: 이산 코사인 변환 기반의 CTPC 모듈)의 곱셈 연산의 수가 적다는 것을 알 수 있다.
- [0055] 즉, 종래의 채널 방향의 점별 컨벌루션 모듈의 연산에서는 상대적으로 많은 수의 곱셈 연산이 요구된 반면, 본원 발명의 이산 코사인 변환 기반 또는 이산 윌시 하다마드 변환 기반의 CTPC 모듈은 상대적으로 적은 수의 곱셈 연산이 요구됨을 알 수 있다. 특히, 선 207을 참조할 때, 이산 윌시 하다마드 변환 기반의 CTPC 모듈의 연산에서는, 곱셈 연산이 요구되지 않고, 덧셈 및 뺄셈 연산만이 요구됨을 알 수 있다.
- [0056] 본원 발명의 CTPC 모듈(201a, 201b)의 연산에 따라서, 종래의 채널 방향의 점별 컨벌루션 모듈(117a, 117b)에 따른 연산이 요구하였던 $O(N^2)$ 의 계산 복잡도에 비해, $O(N \log N)$ 의 계산 복잡도로 감소될 수 있고, 종래의 채널 방향의 점별 컨벌루션 모듈(117a, 117b)에 따른 연산은 $N \times M$ 개의 가중치 파라미터를 요구한 반면, 가중치 파라미터를 요구하지 않으므로, 인공 신경망의 고속화 및 경량화가 가능해질 수 있다. 특히, 이산 윌시-하다마드 변환을 이용한 점별 컨벌루션 연산 과정에서는 부동 소수점 곱셈 연산이 요구되지 않고, $O(N \log N)$ 의 계산 복잡도의 덧셈과 뺄셈만이 요구되므로, 기존의 점별 컨벌루션에 비해 연산 오버헤드(overhead)가 감소되고, 입력 특징 맵 채널 간의 상관 관계가 포착될 수 있다. 이로 인하여, 부동 소수점 곱셈에 최적화된 GPU(Graphics Processing Unit) 자원의 사용에 제한적인 모바일 장치에서 값비싼 GPU 자원에 대한 의존도가 낮아질 수 있고, 부동 소수점 덧셈 및 뺄셈 모듈만이 존재하는 연산 오버헤드가 적게 드는 연산 처리 유닛만이 이용되어도 컨벌루션 연산이 수행될 수 있다. 또한, 본원 발명에 따르면, 학습 가능한 가중치 파라미터가 요구되지 않으므로, 이를 메모리 공간에 저장하거나, 메모리 접근 비용(memory access cost)이 존재하지 않으므로, 메모리 공간을 제한적으로 사용 가능한 임베디드 시스템 장치(embedded system device)에 대하여 효율적으로 적용될 수 있다.
- [0057] 도 3은, 다양한 실시예들에 따른, 이산 윌시-하다마드 변환을 이용한 CTPC 모듈의 동작 방법을 설명하기 위한 알고리즘이다.
- [0058] 다양한 실시예들에 따라서, 입력 특징 맵 X 는 크기가 $B \times N \times H \times W$ 인 4차원의 특성 벡터이고, 입력 특징 맵 개수는 N 이고, 출력 특징 맵 개수는 M 일 수 있다. 각 출력 채널들의 각각의 필드가 N 이라는 것은, 각 출력 채널이 $\log_2 N$ 반복을 통하여 모든 입력 채널들에 대하여 모두(fully) 반영됨을 의미할 수 있다. 이것은, 어떤 채널 요소들이 가산되고 감산되어 결정론적인 구조로 구성되는지의 계산 과정에도 불구하고 입력 채널 상관(correlation)을 포착(capture)할 수 있다는 좋은 특성이 제공될 수 있음을 의미할 수 있다.
- [0059] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, n 이라는 변수에 $\log N$ 을 저장할 수 있다. (line 1)
- [0060] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, 입력 특징 맵 개수(N)보다 출력 특징 맵 개수(M)가 큰지 여부를 판단할 수 있다.
- [0061] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, 입력 특징 맵 개수(N)보다 출력 특징 맵 개수(M)가 크다면, 채널 축에 대하여 입력 특징 맵 X 에 대하여 0값을 패딩(padding)할 수 있다. (line 2-4)
- [0062] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, n 번 반복에서(line 5), 각 반복 마다 짝수번째 인덱스(다른 말로, 색인)를 가지는 채널 요소들 e 와, 홀수번째 인덱스를 가지는 채널 요소들 o 에 대하여, 이를 요소별로 덧셈 연산을 한 값과 뺄셈 연산을 한 값은 각각 앞에서부터 $N/2$ 번째의 채널 요소들과 $N/2$ 번째부터 마지막 번째까

지의 채널 요소들로 대체할 수 있다. (line 6-10)

- [0063] 다양한 실시예들에 따른 CTPC 모듈(201a, 201b)은, 입력 특징 맵 개수(N)보다 출력 특징 맵 개수(M)가 작다면, 채널 요소에 대하여 앞에서부터 출력 특징 맵 개수까지를 제외한 채널 요소들을 절삭할 수 있다.
- [0064] 도 4a는, 다양한 실시예들에 따른, 점별 컨벌루션 모듈을 포함하는 블록 구조들을 설명하기 위한 예시 도면이다.
- [0065] 도 4a의 (a)는 종래의 Ma et al.이 제안한 ShuffleNet-V2의 블록 구조를 도시하고, 도 4a의 (b), (c), (d)는 본원 발명에 따른 블록 구조들을 도시한다.
- [0066] 표 1은, 도 4a의 (a) 내지 (d)의 Top-1 Accuracy(%), Weight의 수(M), FLOP(floating-point operation)의 수(M)의 비교 결과이다.
- [0067] 표 1의 Baseline (a), (b)는 각각 도 4a의 (a), (b)에 대응하고, 표 1의 (c)-DWHT, (c)-DCT는 도 4a의 (c)에 대응하고, 표 1의 (d)-DWHT, (d)-DCT는 도 4a의 (d)에 대응할 수 있다.
- [0068] 여기서, 표 1의 결과들은 모두, width hyper-parameter 1.1x인 ShuffleNet-V2에서, 출력 채널의 수를 128, 256, 512로 하여 진행된 결과이다.

표 1

	Top-1 Acc(%)	# of Weights(M)	# of FLOPs(M)
Baseline (a)	71.71±0.25	1.57	105
(b)	68.16±0.07	1.57	105
(c)-DWHT	65.89±0.26	0.92	52
(c)-DCT	66.55±0.5	0.92	54
(d)-DWHT	69.13±0.083	0.92	52
(d)-DCT	69.23±0.14	0.92	54

- [0070] 도 4a의 (c) 및 (d)에 도시된 블록 구조는 ReLU 포함 여부에서 차이가 있다. 표 1의 (c)-DWHT 및 (c)-DCT과, (d)-DWHT 및 (d)-DCT를 비교할 때, ReLU 활성 함수가 CTPC 모듈을 포함하는 본원 발명에 있어서, 신경망 네트워크의 정확도(Top-1 Accuracy)에 악영향을 준다는 것을 알 수 있다.
- [0071] 도 4a의 (b) 및 (d)에 도시된 블록 구조는, 각각 RCPC 모듈과 CTPC 모듈을 가진다는 것에서 차이가 있다. RCPC(random constant pointwise convolution) 모듈은, $u(-1/\sqrt{N/2}, 1/\sqrt{N/2})$ 의 분포에서 랜덤으로 초기화되어 학습 과정에서 가중치를 고정하는 PC 레이어를 의미할 수 있다. 표 1의 (b)와 (d)-DWHT 및 (d)-DCT를 비교할 때, Top-1 Accuracy, Weight의 수, FLOP의 수 측면에서, 도 4a의 (d)에 도시된 블록 구조가 월등함을 알 수 있다. 이는, 이산 코사인 변환 또는 이산 월시-하다마드 변환 기반의 CTPC 모듈에 의해, 크로스-채널 상관(cross-channel correlation)의 의미 있는 정보가 추출될 수 있음을 의미할 수 있다.
- [0072] 표 1의 Baseline(a)와, (d)-DCT 및 (d)-DWHT를 비교할 때, Top-1 Accuracy는 약 2.5%정도 낮아지는 대신, Weight의 수 및 FLOP의 수는 각각 41% 및 50%만큼 감소되었음을 알 수 있다.
- [0073] 도 4b는, 다양한 실시예들에 따른, 본원 발명의 점별 컨벌루션 모듈이 적용되는 신경망 네트워크 구조에서의 계층 레벨(hierarchy level)에 따른 성능을 비교하기 위한 그래프들이다.
- [0074] 도 4b의 (a) 내지 (f)는, CIFAR 100에 대하여, 종래의 ShuffleNet-V2의 구조(architecture)에 따른 결과(401a, 401b, 401c, 401d, 401e, 401f)와, 종래의 ShuffleNet-V2의 구조(architecture)에서 일부 계층 레벨에 본원 발명의 CTPC 모듈을 적용(대체)한 결과(403a, 403b, 403c, 403d, 403e, 403f, 405a, 405b, 405c, 405d, 405e, 405f)이다. 여기서, Baseline은 종래의 점별 컨벌루션 모듈에 따른, hyper-parameter가 0.5x, 1x, 1.1x, 1.5x 일 때의 결과를 나타내고, DCT on high level layers 및 DWHT on high level layers는 본원 발명의 이산 코사인 변환 및 이산 월시-하다마드 변환 기반의 CTPC 모듈에 따른, hyper-parameter가 1.1x일 때의 3개의 네트워크 인스턴스(network instances)일 때의 결과를 나타낸다.
- [0075] 종래의 ShuffleNet-V2의 구조(architecture)에서, 도 4b의 (a) 및 (d)는 높은 계층 레벨(high-level layers)에 본원 발명의 CTPC 모듈을 적용한 결과(403a, 405a, 403d, 405d)이고, 도 4b의 (b) 및 (e)는 중간 계층 레벨(middle-level layers)에 본원 발명의 CTPC 모듈을 적용한 결과(403b, 405b, 403e, 405e)이고, 도 4b의 (c) 및

(f)는 낮은 계층 레벨(low level layers)에 본원 발명의 CTPC 모듈을 적용한 결과(403c, 405c, 403f, 405f)이다.

[0076] 상술한 도면들을 참조할 때, 본원 발명의 CTPC 모듈을 높은 계층 레벨에 적용한 경우에 가중치 파라미터들 및 FLOPs와 정확도 간의 트레이드 오프(trade-off)가 더 효율적이라는 것을 알 수 있다.

[0077] 또한, 본원 발명의 이산 코사인 변환 및 이산 월시-하다마드 변환 기반의 CTPC 모듈 중에서, 이산 월시-하다마드 변환 기반의 CTPC 모듈의 효율(예: 정확도)이 더 높거나 같다는 것을 알 수 있다. 이는, 이산 월시-하다마드 변환이 곱셈 연산(예: 부동 소수점 곱셈 연산)을 요구하지 않고 더 적은 수의 덧셈 연산 및 뺄셈 연산만을 요구하므로, 크로스-채널(cross-channel) 정보를 추출하는 성능이 더 뛰어나다는 것을 의미한다.

[0078] 도 4b의 (a) 내지 (f)는 종래의 ShuffleNet-V2의 구조(architecture)와 관련된 결과이지만, 종래의 MobileNet-V1(Howard et al., 2017)의 구조에 대하여도 이와 유사한 결과(표 2)가 얻어질 수 있다.

표 2

[0079]

	Top-1 Acc(%)	# of Weights(M)	# of FLOPs(M)
Baseline	67.15±0.3	3.31	94
DWHT-3-H	68.19±0.35	1.47	73
DCT-3-H	68.21±0.19	1.47	74
DWHT-6-H	68.65±0.27	0.68	48
DCT-6-H	67.95±0.53	0.68	49

[0080] 표 2에서, (변환 종류)-(적용된 블록의 수)-(적용된 계층 레벨)을 의미한다. 예를 들어, DWHT-3-L는, 종래의 MobileNet-V1(Howard et al., 2017)의 구조에서, 이산 월시-하다마드 변환(DWHT) 기반의 모듈을, 낮은 계층 레벨(L)에서 3개의 블록(3)에 대하여 적용한 결과를 의미한다.

[0081] 표 2를 참조할 때, 특히, DWHT-6-H에서 Baseline과 비교할 때, Weight의 수 및 FLOP의 수 측면에서 79.1% 및 48.4%로 감소된 반면, 1.49%만큼 정확도가 증가되었음을 알 수 있다.

[0082] 표 2를 통해, 본원 발명의 CTPC 모듈을 높은 계층 레벨(H)에 적용한 경우에 가중치 파라미터들 및 FLOPs와 정확도 간의 트레이드 오프(trade-off)가 더 효율적이라는 것을 알 수 있다.

[0083] 도 5는, 본원 발명에 대한 ReLU 활성화 함수의 영향을 설명하기 위한 그래프이다.

[0084] 도 5의 (a) 및 (b)는, DCT 및 DWHT로부터의 계층 레벨 활성화(activation)들이, 높은 계층 레벨(High level)(501a, 501b), 중간 계층 레벨(Mid level)(503a, 503b) 및 낮은 계층 레벨(Low level)(505a, 505b)에서, 양의 값들과 음의 값들을 거의 같은 비율로 가진다는 것을 나타낸다. 이러한 양의 값들은 크로스-채널 상관 정보(cross-channel correlation information)를 가질 수 있다.

[0085] 이는, 수학식 3의 C_m 및 수학식 4의 H_m^D 이 모두, 상술한 바와 같이, 같은 수의 양(positive)의 파라미터들과 음(negative)의 파라미터들을 가지는 기저 커널들이고, 상기 양의 파라미터들과 음의 파라미터들의 절대 값들의 분포가 거의 동일하기 때문이다. 따라서, 본원 발명의 PC 레이어의 활성화에 ReLU를 적용하는 것은, 음의 파라미터들에 포함된 크로스-채널 상관 정보를 버리는(discard) 결과를 가져올 수 있기 때문에, 정확도를 악화시킬 수 있다.

[0086] 이에 대한 실험적인 결과로, 표 1의 (c)-DWHT 및 (c)-DCT과, (d)-DWHT 및 (d)-DCT를 다시 참조할 때, ReLU 활성화 함수가 CTPC 모듈을 포함하는 본원 발명에 있어서, 신경망 네트워크의 정확도(Top-1 Accuracy)에 악영향을 준다는 것을 알 수 있었다.

[0087] 도 6은, 본원 발명의 높은 정확도를 설명하기 위한 그래프이다.

[0088] 도 6에서 “Baseline”은, 종래의 ShuffleNet-V2의 구조(architecture)에 따른 결과를 의미하고, “DCT-3H”은, 본원 발명의 이산 코사인 변환에 기반한 CTPC 모듈을 높은 계층 레벨의 3개의 블록에 대하여 적용한 결과를 의미하고, “DWHT-3H”는, 본원 발명의 이산 월시-하다마드 변환에 기반한 CTPC 모듈을 높은 계층 레벨의 3개의 블록에 대하여 적용한 결과를 의미한다.

[0089] 도 6을 참조할 때, DWHT-3H 및 DCT-3H에서 마지막 3개의 블록들의 3 x 3 공간 방향의 컨벌루션 가중치들(603,

605)이 Baseline(601)보다 더 적은 0 미만의 값들을 가짐을 알 수 있다. 즉, DWHT-3H 및 DCT-3H에서 마지막 3 개의 블록들의 3 x 3 공간 방향의 컨벌루션 가중치들이 Baseline보다 절대값이 0에서 먼 값들을 더 많이 가짐을 나타낸다. 이로부터, 0이 아닌 학습 가능한 가중치가 기존 변환에서 선호하는 최적의 도메인에 적합하다는 것을 알 수 있다.

- [0090] 결과적으로, 이러한 가중치는, 본원 발명의 학습할 수 없는(non-learnable) 변환(예: 이산 코사인 변환 또는 이산 윌시-하다마드 변환)에 대한 정확도 증가에 기여할 수 있다.
- [0091] 도 7은, 높은 계층 레벨에서 3 x 3 공간 방향의 컨벌루션 가중치들의 영향을 설명하기 위한 그래프이다.
- [0092] 도 7은, 가중치 감소 값들(weight decay values)을 변화시키는 가중치들을 정규화(regulate)하면서 실험한 결과이다.
- [0093] 도 7에서 701은, 이산 코사인 변환 기반의 연산을 적용한 결과이고, 703은, 이산 윌시-하다마드 변환 기반의 연산을 적용한 결과이다.
- [0094] 더 높은 가중치 감소 값들은 높은 계층 레벨에서의 3 x 3 공간 방향의 컨벌루션 가중치들을 더 강하게 정규화한다는 것을 알 수 있다. 이로 인하여, 이러한 가중치 값들의 크기에 대한 더 강한 제약이 가중치의 적극적인 활성화(activation)를 방해하기 때문에, 도 6에서 도시된 바와 같이, 정확도가 낮아질 수 있다.
- [0095] 본 문서에 개시된 다양한 실시예들에 따른 전자 장치는 다양한 형태의 장치가 될 수 있다. 전자 장치는, 예를 들면, 휴대용 통신 장치 (예: 스마트폰), 컴퓨터 장치, 휴대용 멀티미디어 장치, 휴대용 의료 기기, 카메라, 웨어러블 장치, 또는 가전 장치를 포함할 수 있다. 본 문서의 실시예에 따른 전자 장치는 전술한 기기들에 한정되지 않는다.
- [0096] 본 문서의 다양한 실시예들 및 이에 사용된 용어들은 본 문서에 기재된 기술적 특징들을 특정한 실시예들로 한정하려는 것이 아니며, 해당 실시예의 다양한 변경, 균등물, 또는 대체물을 포함하는 것으로 이해되어야 한다. 도면의 설명과 관련하여, 유사한 또는 관련된 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다. 아이템에 대응하는 명사의 단수 형은 관련된 문맥상 명백하게 다르게 지시하지 않는 한, 상기 아이템 한 개 또는 복수 개를 포함할 수 있다. 본 문서에서, "A 또는 B", "A 및 B 중 적어도 하나", "A 또는 B 중 적어도 하나," "A, B 또는 C," "A, B 및 C 중 적어도 하나," 및 "A, B, 또는 C 중 적어도 하나"와 같은 문구들 각각은 그 문구들 중 해당하는 문구에 함께 나열된 항목들 중 어느 하나, 또는 그들의 모든 가능한 조합을 포함할 수 있다. "제 1", "제 2", 또는 "첫째" 또는 "둘째"와 같은 용어들은 단순히 해당 구성요소를 다른 해당 구성요소와 구분하기 위해 사용될 수 있으며, 해당 구성요소들을 다른 측면(예: 중요성 또는 순서)에서 한정하지 않는다. 어떤(예: 제 1) 구성요소가 다른(예: 제 2) 구성요소에, "기능적으로" 또는 "통신적으로" 라는 용어와 함께 또는 이런 용어 없이, "커플드" 또는 "커넥티드" 라고 언급된 경우, 그것은 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로(예: 유선으로), 무선으로, 또는 제 3 구성요소를 통하여 연결될 수 있다는 것을 의미한다.
- [0097] 본 문서에서 사용된 용어 "모듈"은 하드웨어, 소프트웨어 또는 펌웨어로 구현된 유닛을 포함할 수 있으며, 예를 들면, 로직, 논리 블록, 부품, 또는 회로 등의 용어와 상호 호환적으로 사용될 수 있다. 모듈은, 일체로 구성된 부품 또는 하나 또는 그 이상의 기능을 수행하는, 상기 부품의 최소 단위 또는 그 일부가 될 수 있다. 예를 들면, 일실시예에 따르면, 모듈은 ASIC(application-specific integrated circuit)의 형태로 구현될 수 있다.
- [0098] 본 문서의 다양한 실시예들은 기기(machine)(예: 전자 장치) 의해 읽을 수 있는 저장 매체(storage medium)(예: 내장 메모리 또는 외장 메모리)에 저장된 하나 이상의 명령어들을 포함하는 소프트웨어(예: 프로그램)로서 구현될 수 있다. 예를 들면, 기기(예: 전자 장치)의 프로세서는, 저장 매체로부터 저장된 하나 이상의 명령어들 중 적어도 하나의 명령을 호출하고, 그것을 실행할 수 있다. 이것은 기기가 상기 호출된 적어도 하나의 명령어에 따라 적어도 하나의 기능을 수행하도록 운영되는 것을 가능하게 한다. 상기 하나 이상의 명령어들은 컴파일러에 의해 생성된 코드 또는 인터프리터에 의해 실행될 수 있는 코드를 포함할 수 있다. 기기로 읽을 수 있는 저장매체는, 비일시적(non-transitory) 저장매체의 형태로 제공될 수 있다. 여기서, '비일시적'은 저장매체가 실재(tangible)하는 장치이고, 신호(signal)(예: 전자기파)를 포함하지 않는다는 것을 의미할 뿐이며, 이 용어는 데이터가 저장매체에 반영구적으로 저장되는 경우와 임시적으로 저장되는 경우를 구분하지 않는다.
- [0099] 일실시예에 따르면, 본 문서에 개시된 다양한 실시예들에 따른 방법은 컴퓨터 프로그램 제품(computer program product)에 포함되어 제공될 수 있다. 컴퓨터 프로그램 제품은 상품으로서 판매자 및 구매자 간에 거래될 수 있다. 컴퓨터 프로그램 제품은 기기로 읽을 수 있는 저장 매체(예: compact disc read only memory (CD-ROM))의 형태로 배포되거나, 또는 어플리케이션 스토어(예: 플레이 스토어™)를 통해 또는 두개의 사용자 장치들(예: 스

마트폰들) 간에 직접, 온라인으로 배포(예: 다운로드 또는 업로드)될 수 있다. 온라인 배포의 경우에, 컴퓨터 프로그램 제품의 적어도 일부는 제조사의 서버, 어플리케이션 스토어의 서버, 또는 중계 서버의 메모리와 같은 기기로 읽을 수 있는 저장 매체에 적어도 일시 저장되거나, 임시적으로 생성될 수 있다.

[0100] 다양한 실시예들에 따르면, 상기 기술한 구성요소들의 각각의 구성요소(예: 모듈 또는 프로그램)는 단수 또는 복수의 개체를 포함할 수 있다. 다양한 실시예들에 따르면, 전술한 해당 구성요소들 중 하나 이상의 구성요소들 또는 동작들이 생략되거나, 또는 하나 이상의 다른 구성요소들 또는 동작들이 추가될 수 있다. 대체적으로 또는 추가적으로, 복수의 구성요소들(예: 모듈 또는 프로그램)은 하나의 구성요소로 통합될 수 있다. 이런 경우, 통합된 구성요소는 상기 복수의 구성요소들 각각의 구성요소의 하나 이상의 기능들을 상기 통합 이전에 상기 복수의 구성요소들 중 해당 구성요소에 의해 수행되는 것과 동일 또는 유사하게 수행할 수 있다. 다양한 실시예들에 따르면, 모듈, 프로그램 또는 다른 구성요소에 의해 수행되는 동작들은 순차적으로, 병렬적으로, 반복적으로, 또는 휴리스틱하게 실행되거나, 상기 동작들 중 하나 이상이 다른 순서로 실행되거나, 생략되거나, 또는 하나 이상의 다른 동작들이 추가될 수 있다.

부호의 설명

[0101]

115: 채널 분리 모듈

117a, 117b: 채널 방향의 점별 컨벌루션 모듈

119a, 119b, 119c: 배치 단위의 정규화 모듈

121a, 121b: ReLU

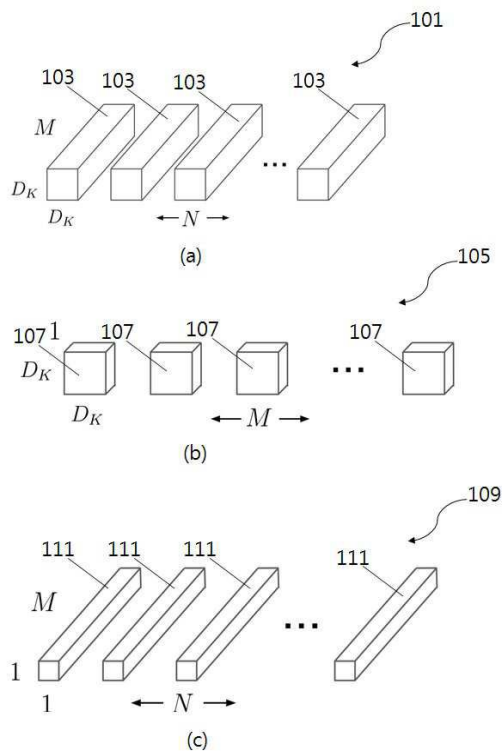
123: 공간 방향의 컨벌루션 모듈

125: 채널 통합 모듈

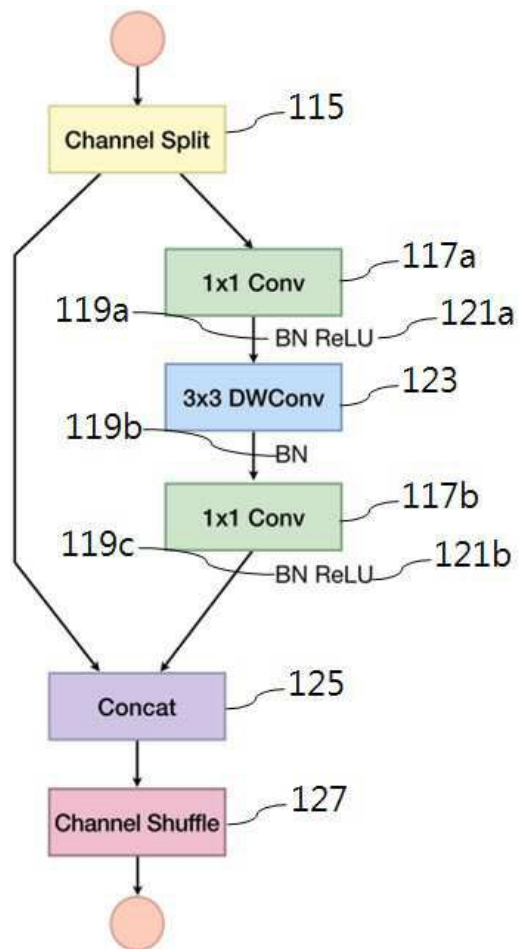
127: 채널 혼합 모듈

도면

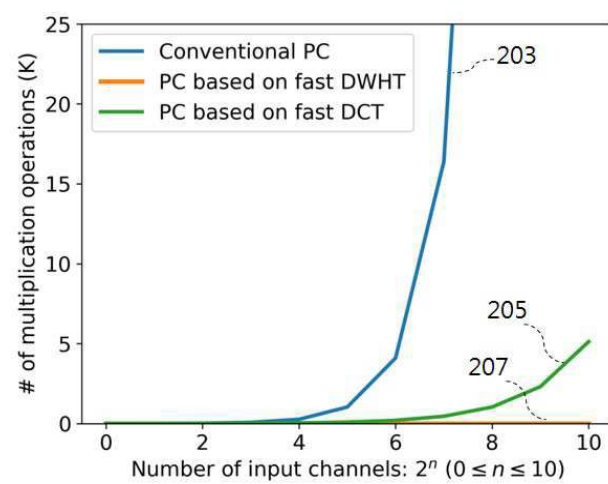
도면1



도면2a



도면2b



도면3

Algorithm 1 A new pointwise convolution using fast DWHT algorithm

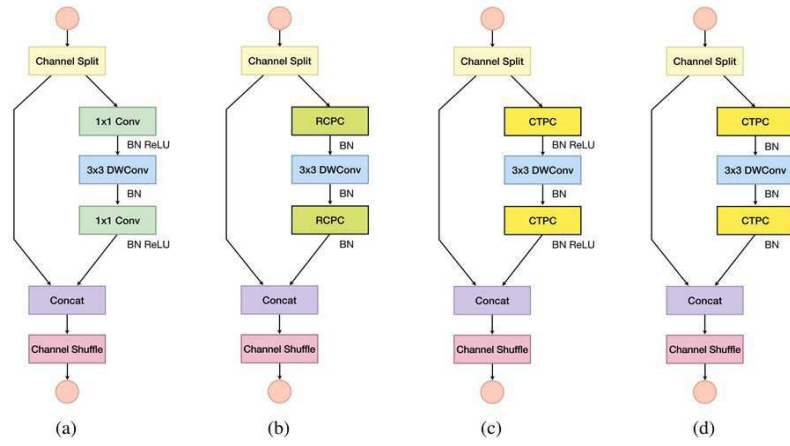
Input: 4D input features $X(B \times N \times H \times W)$, output channel M

```

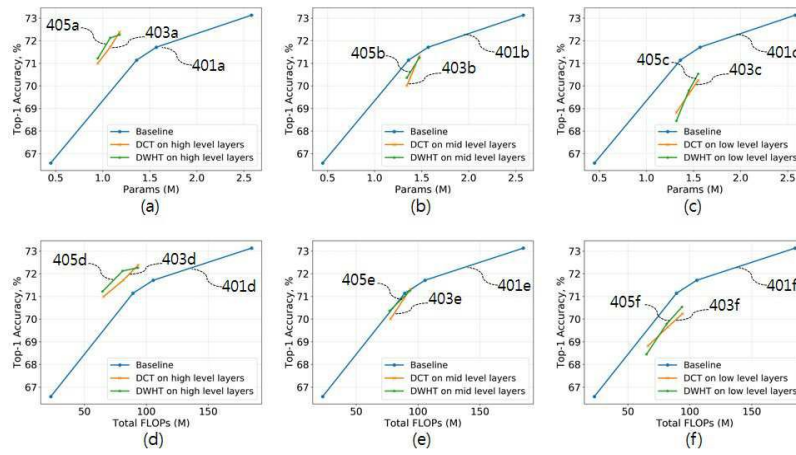
1:  $n \leftarrow \log_2 N$ 
2: if  $N < M$  then
3:   ZeroPad1D( $X$ , axis=1) ▷ pad zeros along the channel axis
4: end if
5: for  $i \leftarrow 1$  to  $n$  do
6:    $e \leftarrow X[:, :, :, 2^i : 2^{i+1}]$ 
7:    $o \leftarrow X[:, :, :, 1 : 2^i]$ 
8:    $X[:, :, :, N/2 : N/2 + 2^i] \leftarrow e + o$ 
9:    $X[:, :, :, N/2 + 2^i : N] \leftarrow e - o$ 
10: end for
11: if  $N > M$  then
12:    $X \leftarrow X[:, :, :, M : N]$ 
13: end if

```

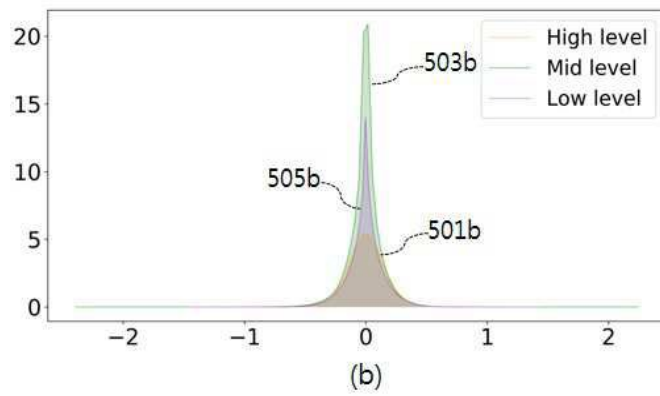
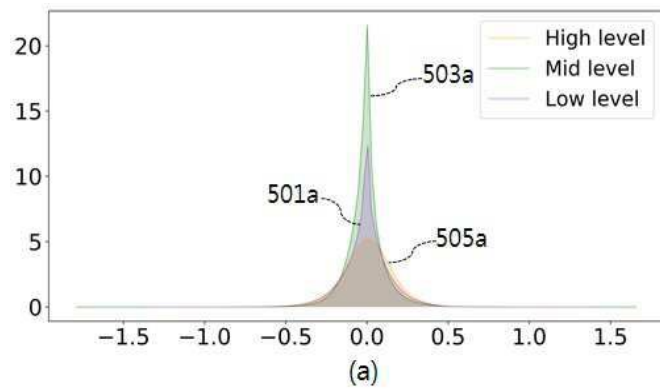
도면4a



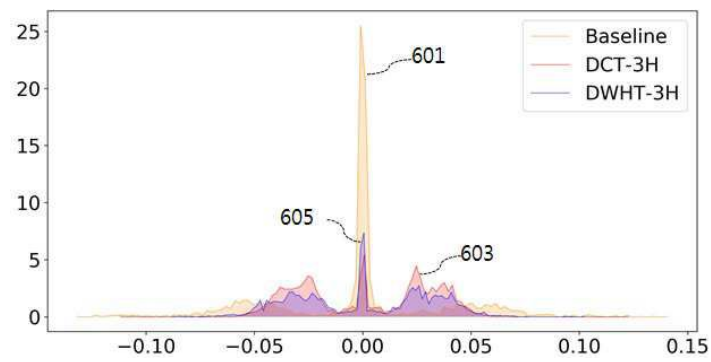
도면4b



도면5



도면6



도면7

