(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0023006
(43) 공개일자 2021년03월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) *G06N 3/08* (2006.01)
H04N 1/413 (2006.01) *H04N 19/00* (2014.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2019-0102554
(22) 출원일자 2019년08월21일
심사청구일자 2019년08월21일

(71) 출원인
네이버웹툰 유한회사
경기도 성남시 분당구 분당내곡로 117, 9층(백현동)
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
장재혁
경기도 성남시 분당구 분당내곡로 117 크래프톤타워 9층
이종석
서울특별시 서대문구 연세로 50, 연세대학교 (신촌동)
(뒷면에 계속)
(74) 대리인
양성보

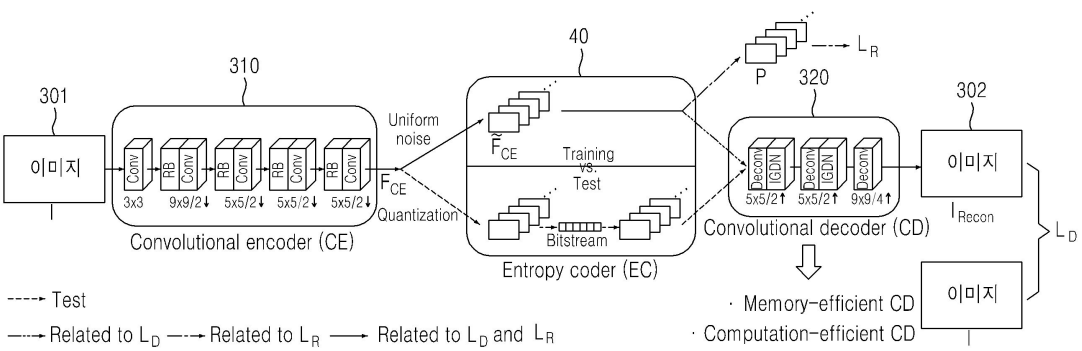
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **딥러닝 기반 이미지 압축 효율 향상을 위한 방법 및 시스템**

(57) 요 약

딥러닝 기반 이미지 압축 효율 향상을 위한 방법 및 시스템이 개시된다. 이미지의 압축(compression)을 위한 딥러닝 네트워크 구조의 인코더를 이미지의 압축 해제(decompression)를 위한 딥러닝 네트워크 구조의 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 구성할 수 있으며, 디코더에 프루닝(pruning)을 적용하여 디코더를 구성하는 딥러닝 네트워크의 적어도 일부 연결을 제거할 수 있다.

대 표 도



(52) CPC특허분류
H04N 1/413 (2013.01)
H04N 19/00 (2013.01)
(72) 발명자
최준호
인천광역시 남동구 논현로26번길 12, 605A(논현동)
김준혁
인천광역시 연수구 송도과학로 85, 기숙사 F동(송도동)

명 세 서

청구범위

청구항 1

컴퓨터 시스템에 있어서,

메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로세서는,

이미지의 압축(compression)을 위한 딥러닝 네트워크 구조의 인코더

를 포함하고,

상기 인코더는 상기 이미지의 압축 해제(decompression)를 위한 딥러닝 네트워크 구조의 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어진 것

을 특징으로 하는 컴퓨터 시스템.

청구항 2

제1항에 있어서,

상기 인코더는 CNN(convolution neural network) 모델로서 콘볼루션 계층(convolution layer)과 GDN 계층(generalized divisible normalization layer) 및 스킵 연결(skip connection)을 가진 비대칭 콘볼루션 오토인코더 구조로 구성되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 3

제1항에 있어서,

상기 인코더는 잔여 블록(residual block) 계층과 콘볼루션 계층의 결합이 반복되는 비대칭 콘볼루션 오토인코더 구조로 구성되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 4

제3항에 있어서,

상기 잔여 블록 계층은 콘볼루션 계층, GDN 계층, 콘볼루션 계층 순의 결합과 입력과 출력 간의 스킵 연결로 구성되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 5

제3항에 있어서,

상기 인코더는 상기 이미지의 추가 정보로서 상기 이미지에서 분석된 주파수 정보가 적어도 하나의 콘볼루션 계층으로 입력되는 구조를 포함하는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 6

제1항에 있어서,

상기 인코더를 구성하는 딥러닝 네트워크의 목적 함수는 수학식 1과 같이 정의되는 것

을 특징으로 하는 컴퓨터 시스템.

[수학식 1]

$$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_D$$

(여기서, $\mathcal{L}_R$ 은 비율(rate) 관련 손실 조건을 나타내고, $\mathcal{L}_D$ 는 왜곡(distortion) 관련 손실 조건을 나타내며, λ 는 $\mathcal{L}_R$ 과 $\mathcal{L}_D$ 를 제어하기 위한 가중치를 나타낸다.)

청구항 7

컴퓨터 시스템에 있어서,

메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로세서는,

인코더로부터 수신된 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더

를 포함하고,

상기 인코더는 상기 이미지의 압축을 위한 딥러닝 네트워크 구조로서 상기 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고,

상기 디코더는 프루닝(pruning)을 통해 상기 디코더를 구성하는 딥러닝 네트워크의 적어도 일부 연결이 제거되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 8

제7항에 있어서,

상기 디코더는 상기 디코더를 구성하는 딥러닝 네트워크의 각 계층에 포함된 필터 별로 해당 필터가 가진 가중치 중 일부 가중치가 제거되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 9

제8항에 있어서,

각 가중치의 크기가 프루닝 기준이 되는 것으로,

상기 디코더를 구성하는 딥러닝 네트워크의 모든 계층을 대상으로 각 계층 별로 프루닝 임계값이 설정되어 각 필터의 가중치 중 상기 프루닝 임계값 미만에 해당되는 가중치가 제거되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 10

제7항에 있어서,

상기 디코더는 상기 디코더를 구성하는 딥러닝 네트워크의 각 계층 별로 해당 계층에 포함된 필터 중 일부 필터가 제거되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 11

제10항에 있어서,

각 필터의 가중치 절대값의 평균이 프루닝 기준이 되는 것으로,

상기 디코더를 구성하는 딥러닝 네트워크에서 마지막 계층을 제외한 나머지 계층을 대상으로 각 계층 별로 프루닝 임계값이 설정되어 해당 계층의 필터 중 가중치 절대값의 평균이 상기 프루닝 임계값을 초과하는 필터가 제거되는 것

을 특징으로 하는 컴퓨터 시스템.

청구항 12

컴퓨터 시스템에서 실행되는 방법에 있어서,

상기 컴퓨터 시스템은 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는 이미지의 압축을 위한 딥러닝 네트워크 구조의 인코더를 포함하고,

상기 인코더는 상기 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고,

상기 방법은,

상기 비대칭 네트워크 구조의 인코더를 통해 상기 이미지를 잠재 표현(latent representation)으로 변환함으로써 압축하는 단계

를 포함하는 방법.

청구항 13

제12항에 있어서,

상기 압축하는 단계는,

상기 이미지의 추가 정보로서 상기 이미지에서 분석된 주파수 정보를 이용하여 상기 이미지를 압축하는 단계

를 포함하는 방법.

청구항 14

컴퓨터 시스템에서 실행되는 방법에 있어서,

상기 컴퓨터 시스템은 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는 인코더로부터 수신된 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더를 포함하고,

상기 인코더는 상기 이미지의 압축을 위한 딥러닝 네트워크 구조로서 상기 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고,

상기 방법은,

상기 인코더로부터 수신된 이미지를 상기 디코더를 통해 재구성하여 복원 이미지를 생성하는 단계

를 포함하는 방법.

청구항 15

제14항에 있어서,

상기 생성하는 단계는,

상기 디코더를 구성하는 딥러닝 네트워크의 적어도 일부 연결을 프루닝하는 단계

를 포함하는 방법.

청구항 16

제14항에 있어서,

상기 생성하는 단계

상기 디코더를 구성하는 딥러닝 네트워크의 각 계층에 포함된 필터 별로 해당 필터가 가진 가중치 중 일부 가중치를 프루닝하는 단계

를 포함하는 방법.

청구항 17

제16항에 있어서,

상기 프루닝하는 단계는,

상기 디코더를 구성하는 딥러닝 네트워크의 모든 계층을 대상으로 각 계층 별로 프루닝 임계값을 설정하여 각 필터의 가중치 중 상기 프루닝 임계값 미만에 해당되는 가중치를 제거하는 것

을 특징으로 하는 방법.

청구항 18

제14항에 있어서,

상기 생성하는 단계

상기 디코더를 구성하는 딥러닝 네트워크의 각 계층 별로 해당 계층에 포함된 필터 중 일부 필터를 프루닝하는 단계

를 포함하는 방법.

청구항 19

제18항에 있어서,

상기 프루닝하는 단계는,

상기 디코더를 구성하는 딥러닝 네트워크에서 마지막 계층을 제외한 나머지 계층을 대상으로 각 계층 별로 프루닝 임계값을 설정하여 해당 계층의 필터 중 가중치 평균이 상기 프루닝 임계값을 초과하는 필터를 제거하는 것

을 특징으로 하는 방법.

청구항 20

제12항 또는 제19항의 방법을 상기 컴퓨터 시스템에 실행시키기 위해 비-일시적인 컴퓨터 판독가능한 기록 매체에 저장되는 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 딥러닝 기반의 이미지 압축 기술에 관한 것이다.

배 경 기 술

[0002] 데이터 압축 알고리즘은 무손실 압축과 손실 압축으로 크게 나누어진다.

[0003] 텍스트를 압축할 경우에는 RLC(Run Length Code)와 허프만 코드(Huffman Code) 등의 무손실 압축 알고리즘을 사용한다. 반면, 이미지를 압축할 경우에는 JPEG와 MPEG 등의 손실 압축 알고리즘을 사용한다.

[0004] 손실 이미지 압축은 불가피한 품질 저하를 용인하면서 이미지에 있는 정보의 일부를 폐기함으로써 가능한 한 적

은 용량(비율)의 영상을 인코딩하는 작업을 의미한다. 대용량 이미지의 급속한 발전과 대중화로 인해 핵심적 이미지 처리 문제는 더욱더 필수적이다.

[0005] 이미지 압축 기술의 일례로서, 한국 공개특허 제10-2009-0050325호(공개일 2009년 05월 20일)에는 데이터가 존재하는 유효 라인들을 공통요소를 갖는 라인별로 그룹화하여 그룹별로 유효 라인들의 데이터를 압축하고 암호화하는 기술이 개시되어 있다.

발명의 내용

해결하려는 과제

[0006] 딥러닝 기반 이미지 압축 기술로서 비대칭 오토인코더 아키텍처와 디코더 프루닝(pruning)에 기반한 효율적인 손실 이미지 압축 기술을 제공한다.

[0007] 콘볼루션 계층(convolutional layer), GDN 계층(generalized divisible normalization layer), 스킵 연결(skip connection)을 가진 비대칭적 콘볼루션 오토인코더 구조를 제공한다.

[0008] 가중치 프루닝(weight pruning) 또는 필터 프루닝(filter pruning)을 적용하여 보다 가볍고 빠른 디코더 구조를 제공한다.

과제의 해결 수단

[0009] 컴퓨터 시스템에 있어서, 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 이미지의 압축(compression)을 위한 딥러닝 네트워크 구조의 인코더를 포함하고, 상기 인코더는 상기 이미지의 압축 해제(decompression)를 위한 딥러닝 네트워크 구조의 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어진 것을 특징으로 하는 컴퓨터 시스템을 제공한다.

[0010] 일 측면에 따르면, 상기 인코더는 CNN(convolution neural network) 모델로서 콘볼루션 계층(convolution layer)과 GDN 계층(generalized divisible normalization layer) 및 스킵 연결(skip connection)을 가진 비대칭 콘볼루션 오토인코더 구조로 구성될 수 있다.

[0011] 다른 측면에 따르면, 상기 인코더는 잔여 블록(residual block) 계층과 콘볼루션 계층의 결합이 반복되는 비대칭 콘볼루션 오토인코더 구조로 구성될 수 있다.

[0012] 또 다른 측면에 따르면, 상기 잔여 블록 계층은 콘볼루션 계층, GDN 계층, 콘볼루션 계층 순의 결합과 입력과 출력 간의 스킵 연결로 구성될 수 있다.

[0013] 또 다른 측면에 따르면, 상기 인코더는 상기 이미지의 추가 정보로서 상기 이미지에서 분석된 주파수 정보가 적어도 하나의 콘볼루션 계층으로 입력되는 구조를 포함할 수 있다.

[0014] 또 다른 측면에 따르면, 상기 인코더를 구성하는 딥러닝 네트워크의 목적 함수는 수학식 1과 같이 정의된다.

[0015] [수학식 1]

[0016] $$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_D$$

[0017] (여기서, $\mathcal{L}_R$ 은 비율(rate) 관련 손실 조건을 나타내고, $\mathcal{L}_D$ 는 왜곡(distortion) 관련 손실 조건을 나타내며, λ 는 $\mathcal{L}_R$ 과 $\mathcal{L}_D$ 를 제어하기 위한 가중치를 나타낸다.)

[0018] 컴퓨터 시스템에 있어서, 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 인코더로부터 수신된 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더를 포함하고, 상기 인코더는 상기 이미지의 압축을 위한 딥러닝 네트워크 구조로서 상기 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고, 상기 디코더는 프루닝(pruning)을 통해 상기 디코더를 구성하는 딥러닝 네트워크의 적어도 일부 연결이 제거되는 것을 특징으로 하는 컴퓨터 시스템을 제공한다.

[0019] 일 측면에 따르면, 상기 디코더는 상기 디코더를 구성하는 딥러닝 네트워크의 각 계층에 포함된 필터 별로 해당

필터가 가진 가중치 중 일부 가중치가 제거될 수 있다.

[0020] 다른 측면에 따르면, 각 가중치의 크기가 프루닝 기준이 되는 것으로, 상기 디코더를 구성하는 딥러닝 네트워크의 모든 계층을 대상으로 각 계층 별로 프루닝 임계값이 설정되어 각 필터의 가중치 중 상기 프루닝 임계값 미만에 해당되는 가중치가 제거될 수 있다.

[0021] 또 다른 측면에 따르면, 상기 디코더는 상기 디코더를 구성하는 딥러닝 네트워크의 각 계층 별로 해당 계층에 포함된 필터 중 일부 필터가 제거될 수 있다.

[0022] 또 다른 측면에 따르면, 각 필터의 가중치 절대값의 평균이 프루닝 기준이 되는 것으로, 상기 디코더를 구성하는 딥러닝 네트워크에서 마지막 계층을 제외한 나머지 계층을 대상으로 각 계층 별로 프루닝 임계값이 설정되어 해당 계층의 필터 중 가중치 절대값의 평균이 상기 프루닝 임계값을 초과하는 필터가 제거될 수 있다.

[0023] 컴퓨터 시스템에서 실행되는 방법에 있어서, 상기 컴퓨터 시스템은 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는 이미지의 압축을 위한 딥러닝 네트워크 구조의 인코더를 포함하고, 상기 인코더는 상기 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고, 상기 방법은, 상기 비대칭 네트워크 구조의 인코더를 통해 상기 이미지를 잠재 표현(latent representation)으로 변환함으로써 압축하는 단계를 포함하는 방법을 제공한다.

[0024] 컴퓨터 시스템에서 실행되는 방법에 있어서, 상기 컴퓨터 시스템은 메모리에 포함된 컴퓨터 판독가능한 명령들을 실행하도록 구성된 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는 인코더로부터 수신된 이미지의 압축 해제를 위한 딥러닝 네트워크 구조의 디코더를 포함하고, 상기 인코더는 상기 이미지의 압축을 위한 딥러닝 네트워크 구조로서 상기 디코더보다 많은 개수의 계층을 가진 비대칭 네트워크 구조로 이루어지고, 상기 방법은, 상기 인코더로부터 수신된 이미지를 상기 디코더를 통해 재구성하여 복원 이미지를 생성하는 단계를 포함하는 방법을 제공한다.

[0025] 상기 방법을 상기 컴퓨터 시스템에 실행시키기 위해 비-일시적인 컴퓨터 판독가능한 기록 매체에 저장되는 컴퓨터 프로그램을 제공한다.

발명의 효과

[0026] 본 발명의 실시예들에 따르면, 딥러닝 기반 이미지 압축 기술에서 디코더에 비해 더 깊은 네트워크로 설계된 인코더의 비대칭 구조를 통해 이미지의 비율 왜곡(rate distortion) 성능을 향상시킬 수 있다.

[0027] 본 발명의 실시예들에 따르면, 이미지 압축 기술에서 디코더에 프루닝 기법을 적용함으로써 모델 사이즈 측면에서 더욱 가볍고 처리 시간과 속도 측면에서 보다 빠른 디코더를 제공할 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 일실시예에 따른 네트워크 환경의 예를 도시한 도면이다.

도 2는 본 발명의 일실시예에 있어서 전자 기기 및 서버의 내부 구성을 설명하기 위한 블록도이다.

도 3은 본 발명의 일실시예에 있어서 이미지 압축을 위한 인코더와 이미지 압축 해제를 위한 디코더를 설명하기 위한 예시 도면이다.

도 4는 본 발명의 일실시예에 있어서 이미지 압축(인코딩) 및 압축 해제(디코딩) 위한 심층 네트워크의 전체 아키텍처를 도시한 것이다.

도 5는 본 발명의 일실시예에 있어서 인코더 모델에 포함된 잔여 블록(GB) 계층의 구조를 설명하기 위한 예시 도면이다.

도 6은 본 발명의 일실시예에 있어서 디코더보다 더 깊은 네트워크 구조의 인코더를 설명하기 위한 예시 도면이다.

도 7은 본 발명의 일실시예에 있어서 이미지 압축을 위한 추가 정보로 주파수 정보가 입력되는 네트워크 구조를 설명하기 위한 예시 도면이다.

도 8은 본 발명의 일실시예에 있어서 가중치 프루닝 과정을 설명하기 위한 예시 도면이다.

도 9는 본 발명의 일실시예에 있어서 필터 프루닝 과정을 설명하기 위한 예시 도면이다.

발명을 실시하기 위한 구체적인 내용

[0029] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0031] 본 발명의 실시예들은 딥러닝 기반의 이미지 압축 기술에 관한 것이다.

[0032] 본 명세서에서 구체적으로 개시되는 것들을 포함하는 실시예들은 비대칭 오토인코더 아키텍처와 디코더 프루닝에 기반한 효율적인 손실 이미지 압축 기술을 제공할 수 있고, 이를 통해 이미지 비율 왜곡 성능을 최적화할 수 있고 프루닝 효과에 따라 경량 디코더와 고속 디코더를 구현할 수 있다.

[0033] 도 1은 본 발명의 일실시예에 따른 네트워크 환경의 예를 도시한 도면이다. 도 1의 네트워크 환경은 복수의 전자 기기들(110, 120, 130, 140), 복수의 서버들(150, 160) 및 네트워크(170)를 포함하는 예를 나타내고 있다. 이러한 도 1은 발명의 설명을 위한 일례로 전자 기기의 수나 서버의 수가 도 1과 같이 한정되는 것은 아니다.

[0034] 복수의 전자 기기들(110, 120, 130, 140)은 컴퓨터 시스템으로 구현되는 고정형 단말이거나 이동형 단말일 수 있다. 복수의 전자 기기들(110, 120, 130, 140)의 예를 들면, 스마트폰(smart phone), 휴대폰, 내비게이션, 컴퓨터, 노트북, 디지털방송용 단말, PDA(Personal Digital Assistants), PMP(Portable Multimedia Player), 태블릿 PC, 게임 콘솔(game console), 웨어러블 디바이스(wearable device), IoT(internet of things) 디바이스, VR(virtual reality) 디바이스, AR(augmented reality) 디바이스 등이 있다. 일례로 도 1에서는 전자 기기(110)의 예로 스마트폰의 형상을 나타내고 있으나, 본 발명의 실시예들에서 전자 기기(110)는 실질적으로 무선 또는 유선 통신 방식을 이용하여 네트워크(170)를 통해 다른 전자 기기들(120, 130, 140) 및/또는 서버(150, 160)와 통신할 수 있는 다양한 물리적인 컴퓨터 시스템들 중 하나를 의미할 수 있다.

[0035] 통신 방식은 제한되지 않으며, 네트워크(170)가 포함할 수 있는 통신망(일례로, 이동통신망, 유선 인터넷, 무선 인터넷, 방송망, 위성망 등)을 활용하는 통신 방식뿐만 아니라 기기들간의 근거리 무선 통신 역시 포함될 수 있다. 예를 들어, 네트워크(170)는, PAN(personal area network), LAN(local area network), CAN(campus area network), MAN(metropolitan area network), WAN(wide area network), BBN(broadband network), 인터넷 등의 네트워크 중 하나 이상의 임의의 네트워크를 포함할 수 있다. 또한, 네트워크(170)는 버스 네트워크, 스타 네트워크, 링 네트워크, 메쉬 네트워크, 스타-버스 네트워크, 트리 또는 계층적(hierarchical) 네트워크 등을 포함하는 네트워크 토폴로지 중 임의의 하나 이상을 포함할 수 있으나, 이에 제한되지 않는다.

[0036] 서버(150, 160) 각각은 복수의 전자 기기들(110, 120, 130, 140)과 네트워크(170)를 통해 통신하여 명령, 코드, 파일, 컨텐츠, 서비스 등을 제공하는 컴퓨터 장치 또는 복수의 컴퓨터 장치들로 구현될 수 있다. 예를 들어, 서버(150)는 네트워크(170)를 통해 접속한 복수의 전자 기기들(110, 120, 130, 140)로 제1 서비스를 제공하는 시스템일 수 있으며, 서버(160) 역시 네트워크(170)를 통해 접속한 복수의 전자 기기들(110, 120, 130, 140)로 제2 서비스를 제공하는 시스템일 수 있다. 보다 구체적인 예로, 서버(150)는 복수의 전자 기기들(110, 120, 130, 140)에 설치되어 구동되는 컴퓨터 프로그램으로서의 어플리케이션을 통해, 해당 어플리케이션이 목적하는 서비스(일례로, 웹툰 서비스 등)를 제1 서비스로서 복수의 전자 기기들(110, 120, 130, 140)로 제공할 수 있다. 다른 예로, 서버(160)는 상술한 어플리케이션의 설치 및 구동을 위한 파일을 복수의 전자 기기들(110, 120, 130, 140)로 배포하는 서비스를 제2 서비스로서 제공할 수 있다.

[0037] 도 2는 본 발명의 일실시예에 있어서 전자 기기 및 서버의 내부 구성을 설명하기 위한 블록도이다. 도 2에서는 전자 기기에 대한 예로서 전자 기기(110), 그리고 서버(150)의 내부 구성을 설명한다. 또한, 다른 전자 기기들(120, 130, 140)이나 서버(160) 역시 상술한 전자 기기(110) 또는 서버(150)와 동일한 또는 유사한 내부 구성을 가질 수 있다.

[0038] 전자 기기(110)와 서버(150)는 메모리(211, 221), 프로세서(212, 222), 통신 모듈(213, 223) 그리고 입출력 인터페이스(214, 224)를 포함할 수 있다. 메모리(211, 221)는 비-일시적인 컴퓨터 판독 가능한 기록매체로서, RAM(random access memory), ROM(read only memory), 디스크 드라이브, SSD(solid state drive), 플래시 메모리(flash memory) 등과 같은 비소멸성 대용량 저장 장치(permanent mass storage device)를 포함할 수 있다. 여기서 ROM, SSD, 플래시 메모리, 디스크 드라이브 등과 같은 비소멸성 대용량 저장 장치는 메모리(211, 221)와는 구분되는 별도의 영구 저장 장치로서 전자 기기(110)나 서버(150)에 포함될 수도 있다. 또한, 메모리(211, 221)에는 운영체제와 적어도 하나의 프로그램 코드(일례로 전자 기기(110)에 설치되어 구동되는 브라우저나 특

정 서비스의 제공을 위해 전자 기기(110)에 설치된 어플리케이션 등을 위한 코드)가 저장될 수 있다. 이러한 소프트웨어 구성요소들은 메모리(211, 221)와는 별도의 컴퓨터에서 판독 가능한 기록매체로부터 로딩될 수 있다. 이러한 별도의 컴퓨터에서 판독 가능한 기록매체는 플로피 드라이브, 디스크, 테이프, DVD/CD-ROM 드라이브, 메모리 카드 등의 컴퓨터에서 판독 가능한 기록매체를 포함할 수 있다. 다른 실시예에서 소프트웨어 구성요소들은 컴퓨터에서 판독 가능한 기록매체가 아닌 통신 모듈(213, 223)을 통해 메모리(211, 221)에 로딩될 수도 있다. 예를 들어, 적어도 하나의 프로그램은 개발자들 또는 어플리케이션의 설치 파일을 배포하는 파일 배포 시스템(일례로, 상술한 서버(160))이 네트워크(170)를 통해 제공하는 파일들에 의해 설치되는 컴퓨터 프로그램(일례로 상술한 어플리케이션)에 기반하여 메모리(211, 221)에 로딩될 수 있다.

[0039] 프로세서(212, 222)는 기본적인 산술, 로직 및 입출력 연산을 수행함으로써, 컴퓨터 프로그램의 명령을 처리하도록 구성될 수 있다. 명령은 메모리(211, 221) 또는 통신 모듈(213, 223)에 의해 프로세서(212, 222)로 제공될 수 있다. 예를 들어 프로세서(212, 222)는 메모리(211, 221)와 같은 기록 장치에 저장된 프로그램 코드에 따라 수신되는 명령을 실행하도록 구성될 수 있다.

[0040] 통신 모듈(213, 223)은 네트워크(170)를 통해 전자 기기(110)와 서버(150)가 서로 통신하기 위한 기능을 제공할 수 있으며, 전자 기기(110) 및/또는 서버(150)가 다른 전자 기기(일례로 전자 기기(120)) 또는 다른 서버(일례로 서버(160))와 통신하기 위한 기능을 제공할 수 있다. 일례로, 전자 기기(110)의 프로세서(212)가 메모리(211)와 같은 기록 장치에 저장된 프로그램 코드에 따라 생성한 요청이 통신 모듈(213)의 제어에 따라 네트워크(170)를 통해 서버(150)로 전달될 수 있다. 역으로, 서버(150)의 프로세서(222)의 제어에 따라 제공되는 제어 신호나 명령, 컨텐츠, 파일 등이 통신 모듈(223)과 네트워크(170)를 거쳐 전자 기기(110)의 통신 모듈(213)을 통해 전자 기기(110)로 수신될 수 있다. 예를 들어 통신 모듈(213)을 통해 수신된 서버(150)의 제어 신호나 명령, 컨텐츠, 파일 등은 프로세서(212)나 메모리(211)로 전달될 수 있고, 컨텐츠나 파일 등은 전자 기기(110)가 더 포함할 수 있는 저장 매체(상술한 영구 저장 장치)로 저장될 수 있다.

[0041] 입출력 인터페이스(214)는 입출력 장치(215)와의 인터페이스를 위한 수단일 수 있다. 예를 들어, 입력 장치는 키보드, 마우스, 마이크로폰, 카메라 등의 장치를, 그리고 출력 장치는 디스플레이, 스피커, 햅틱 피드백 디바이스(haptic feedback device) 등과 같은 장치를 포함할 수 있다. 다른 예로 입출력 인터페이스(214)는 터치스크린과 같이 입력과 출력을 위한 기능이 하나로 통합된 장치와의 인터페이스를 위한 수단일 수도 있다. 입출력 장치(215)는 전자 기기(110)와 하나의 장치로 구성될 수도 있다. 또한, 서버(150)의 입출력 인터페이스(224)는 서버(150)와 연결되거나 서버(150)가 포함할 수 있는 입력 또는 출력을 위한 장치(미도시)와의 인터페이스를 위한 수단일 수 있다. 보다 구체적인 예로, 전자 기기(110)의 프로세서(212)가 메모리(211)에 로딩된 컴퓨터 프로그램의 명령을 처리함에 있어서 서버(150)나 전자 기기(120)가 제공하는 데이터를 이용하여 구성되는 서비스 화면이나 컨텐츠가 입출력 인터페이스(214)를 통해 디스플레이에 표시될 수 있다.

[0042] 또한, 다른 실시예들에서 전자 기기(110) 및 서버(150)는 도 2의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 전자 기기(110)는 상술한 입출력 장치(215) 중 적어도 일부를 포함하도록 구현되거나 또는 트랜시버(transceiver), GPS(Global Positioning System) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수도 있다. 보다 구체적인 예로, 전자 기기(110)가 스마트폰인 경우, 일반적으로 스마트폰이 포함하고 있는 가속도 센서나 자이로 센서, 카메라 모듈, 각종 물리적인 버튼, 터치패널을 이용한 버튼, 입출력 포트, 진동을 위한 진동기 등의 다양한 구성요소들이 전자 기기(110)에 더 포함되도록 구현될 수 있다.

[0043] 이하에서는 딥러닝 기반 이미지 압축 효율 향상을 위한 방법 및 시스템의 구체적인 실시예를 설명하기로 한다.

[0044] 도 3은 본 발명의 일실시예에 있어서 이미지 압축을 위한 인코더와 이미지 압축 해제를 위한 디코더를 설명하기 위한 예시 도면이다.

[0045] 딥러닝 기반 손실 이미지 압축(인코딩)-압축해제(디코딩) 방법은 일반적으로 콘볼루션 오토인코더를 기반으로 한다. 이는 웹툰과 같이 가공된 이미지를 다수의 사용자에게 제공하기 위한 기술 분야에서 적용될 수 있다.

[0046] 도 3을 참조하면, 인코더(310)는 콘볼루션 계층을 가진 콘볼루션 인코더(convolutional encoder, CE)로서, 인코딩을 통해 원본 이미지(301)를 잠재 표현(latent representation)으로 변환하며, 이는 엔트로피 코더(Entropy coder, EC)(미도시)를 통해 비트스트림(bitstream)으로 압축된다.

[0047] 디코더(320)는 콘볼루션 계층을 가진 콘볼루션 디코더(convolutional decoder, CD)로서, 인코더(310)로부터 수신된 비트스트림을 잠재 표현으로 변환한 후 그것을 이용하여 이미지를 재구성함으로써 복원 이미지(302)를 생

성한다.

[0048] 인코더(310)와 디코더(320) 간의 이미지 손실은 원본 이미지(301)와 복원 이미지(302) 간의 차이로 정의될 수 있다.

[0049] 일례로, 이미지 압축을 위한 인코더(310)는 서버(150)에 해당되는 컴퓨터 시스템 상에 구현될 수 있고, 이미지 압축 해제를 위한 디코더(320)는 전자 기기(110)에 해당되는 컴퓨터 시스템 상에 구현될 수 있다. 이와 같이, 디코더(320)는 주로 모바일 기기나 임베디드 시스템 등 메모리 및 계산 자원이 한정된 컴퓨터 시스템에 배치되는 경우가 대부분이므로, 디코더(320)의 이미지 처리 효율은 매우 중요한 문제가 된다.

[0050] 본 발명은 상대적으로 낮은 메모리와 계산 복잡성을 요구하면서 동시에 비율 왜곡 성능을 보장할 수 있는 효율적인 이미지 디코더를 제공할 수 있다.

[0051] 첫째, 본 발명에서는 콘볼루션 계층(convolutional layer), GDN 계층(generalized divisible normalization layer), 스킵 연결(skip connection)을 가진 비대칭적 콘볼루션 오토인코더 구조를 제공할 수 있다. 기존에는 대부분 대칭적인 오토인코더를 기반으로 하는데, 인코더는 비율과 왜곡에 모두 관여하나, 디코더의 경우 왜곡에만 관여하기 때문에 비율과 왜곡을 모두 최소화하는데 한계가 있다. 따라서, 본 발명의 실시예에서는 디코더보다 더 깊은 네트워크로 설계된 인코더를 통해 비대칭적인 역할을 모델링하고 비대칭적 콘볼루션 오토인코더 구조를 통해 비율 왜곡 성능을 향상시킬 수 있다.

[0052] 둘째, 본 발명에서는 가중치 프루닝(weight pruning)과 필터 프루닝(filter pruning)을 적용하여 보다 가볍고 빠른 디코더 구조를 제공할 수 있다. 주로 이미지 분류(image classification) 기술 분야에서 사용되는 프루닝 기법을 이미지 압축 해제를 위한 디코더에 사용함으로써 클라이언트 측의 리소스나 연산 부담을 효과적으로 줄일 수 있다.

[0053] 본 발명에 따른 이미지 압축(인코딩)-압축해제(디코딩) 방법은 자연 이미지뿐만 아니라 컴퓨터로 만들어진 멀티미디어 이미지(예컨대, 웹툰 등)에 모두 적용 가능하다.

[0054] 딥러닝 기반 이미지 압축 기술에서는 (1) 주어진 비율에 대한 왜곡 최적화 문제와 (2) 공동 비율 왜곡(Joint rate-distortion) 최적화 문제를 다룰 수 있다.

[0055] (1) 주어진 비율에 대한 왜곡 최적화 문제는 RNN(Recurrent Neural Network) 기반 아키텍처를 이용한 인코딩을 통해 가변 압축 비율을 가능하게 한다. 해당 문제와 관련하여, RNN을 기반으로 이미지의 가변비율을 압축하는 기술, 전해상도 영상을 위해 내용 기반 잔류 스케일링 및 게이트 반복 단위의 변동을 사용하는 기술, 가중된 손실 함수, 히든 상태 프라이밍을 통한 향상된 아키텍처 및 공간 적응적 비트 할당을 통한 사후 처리 기술 등이 연구되고 있다. 기존 방법은 단일 모델로 가변 압축 비율을 제공할 수 있을 뿐, 이미지 압축에 있어 비율과 왜곡을 공동으로 최적화하지는 못한다.

[0056] (2) 공동 비율 왜곡 최적화 문제를 해결하기 위해 E2E(end-to-end) 이미지 압축 방법을 이용한다. E2E 이미지 압축 방법에서는 연속적인 완화(relaxation)를 위해 양자화를 부가 균일 잡음(additive uniform noise)으로 대체한다. 양자화 자체에 근접하는 방법 이외에, 포워드 패스(forward pass)에서 확률적 반올림 연산(stochastic rounding operation)을 사용하고 미분계수(derivative)를 역전사 방향에서 기대치의 미분계수로 대체하는 방법을 이용할 수 있다. 이러한 방법들은 완전히 인수분해된 엔트로피 모델을 사용할 수 있으나, 개선된 엔트로피 추정을 위해 잠재 표현의 맥락을 고려하고 있지 못한다.

[0057] 그리고, 본 발명에서 이미지 압축 기술에 적용하고자 하는 네트워크 프루닝은 신경망의 덜 중요한 연결을 제거하는 것을 목표로 하며, 가중치 프루닝과 필터 프루닝으로 나눌 수 있다.

[0058] 가중치 프루닝은 각 가중치를 개별적으로 제거하는 방식으로, 최대 저장 효율성을 위해 원하는 만큼의 매개변수를 제거할 수 있다. 한편, 필터 프루닝은 선택된 필터를 제거하는 것으로, 프루닝의 정도에 따라 직접적으로 계산 복잡성을 감소시킬 수 있는 이점이 있다.

[0059] 도 4는 본 발명의 일실시예에 있어서 이미지 압축(인코딩) 및 압축 해제(디코딩) 위한 심층 네트워크의 전체 아키텍처를 도시한 것이다.

[0060] 전체 시스템은 콘볼루션 인코더(CE)(310), 엔트로피 코더(EC)(40), 및 콘볼루션 디코더(CD)(320)로 구성될 수 있다.

[0061] 설명의 편의를 위해 콘볼루션 인코더(CE)(310), 엔트로피 코더(EC)(40), 및 콘볼루션 디코더(CD)(320)를 하나의

도면으로 도시하였으나, 바람직하게는 분산된 시스템으로 콘볼루션 인코더(CE)(310)와 엔트로피 코더(EC)(40)는 서버(150)에 해당되는 컴퓨터 시스템 상에 구현되고, 콘볼루션 디코더(CD)(320)는 전자 기기(110)에 해당되는 컴퓨터 시스템 상에 구현된다.

[0062] 원본 이미지(301)는 높이가 H이고 너비가 W라 할 때 $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ 와 같이 정의할 수 있다.

[0063] 콘볼루션 인코더(CE)(310)는 피처 맵(feature map) $\mathbf{F}_{CE} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ (여기서, C는 채널 수)을 추출한다. 콘볼루션 인코더(CE)(310)에서 추출된 피처 맵 $\mathbf{F}_{CE}$ 은 수학식 1과 같다.

[0064] [수학식 1]

[0065] $$\mathbf{F}_{CE} = f_{CE}(\mathbf{I})$$

[0066] 여기서, $f_{CE}(\cdot)$ 는 콘볼루션 인코더(CE)(310)의 함수로서 콘볼루션 계층과 GDN 계층으로 비선형 변환을 수행한다.

[0067] 엔트로피 코더(EC)(40)는 학습 과정과 테스트 과정에 따라 그 역할이 달라지며, 학습 과정 중에는 피처 맵 $\mathbf{F}_{CE}$ 를 직접 양자화하는 대신 부가 균일 잡음(additive uniform noise)을 사용하여 연속적인 완화 과정을 수행한다. 이 과정은 콘볼루션 디코더(CD)(320)의 입력으로 사용되는 $\tilde{\mathbf{F}}_{CE}$ 을 산출한다. 그리고, 엔트로피 코더(EC)(40)는 확률 값 $\mathbf{P} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ 을 예측하는데, 이때 확률 값의 각 요소는 $\tilde{\mathbf{F}}_{CE}$의 요소에 해당한다. 각 채널에 대한 확률 질량 함수는 예측 함수 값 $\mathbf{P}$ 에 근거하여 업데이트될 수 있다.

[0068] 테스트 과정 중 이미지를 인코딩하기 위해 엔트로피 코더(EC)(40)는 피처 맵 $\mathbf{F}_{CE}$를 양자화한 다음 훈련된 확률 질량 함수를 활용하는 범위 코더(range coder)를 사용하여 비트스트림을 생성할 수 있다.

[0069] 마지막으로, 콘볼루션 디코더(CD)(320)는 엔트로피 코더(EC)(40)에서 출력된 피처 맵 $\tilde{\mathbf{F}}_{CE}$에서 재구성된 이미지 $\mathbf{I}_{Recon}$ 를 생성한다.

[0070] 이미지 압축(인코딩) 및 압축 해제(디코딩) 위한 심층 네트워크의 목적 함수는 수학식 2와 같이 정의되는 비율과 왜곡의 공동 최적화를 목표로 한다.

[0071] [수학식 2]

[0072] $$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_D$$

[0073] 여기서, $\mathcal{L}_R$ 은 비율 관련 손실 조건을 나타내고, $\mathcal{L}_D$ 는 왜곡 관련 손실 조건을 나타내며, λ는 두 조건 $\mathcal{L}_R$ 과 $\mathcal{L}_D$ 를 제어하기 위한 가중치를 나타낸다. $\mathcal{L}_R$ 은 확률 값 $\mathbf{P}$ 를 이용한 픽셀당 예측 비트(bpp)로 정의할 수 있고, $\mathcal{L}_D$ 에 대해 MSE(평균 제곱 오차) 손실 함수를 활용할 수 있다.

[0074] 비대칭 콘볼루션 오토인코더

[0075] 일반적인 딥러닝 기반 이미지 압축 기술은 대칭적인 콘볼루션 오토인코더에 기초한다. 이러한 대칭 구조는 필요성에 의해 채택된 것이 아니라 적층 오토인코더를 훈련하는데 사용되는 구조가 유지되는 것에 불과하다.

[0076] 콘볼루션 디코더(CD)(320)는 왜곡과 관련된 손실 조건만을 고려하여 최적화되나, 콘볼루션 인코더(CE)(310)는 왜곡뿐만 아니라 비율도 함께 고려하기 때문에 더 많은 역할을 학습하기 위해 훈련 가능한 매개변수가 더 많이 필요하다.

[0077] 이러한 문제를 해결하기 위해 본 발명에서는 콘볼루션 인코더(CE)(310)를 콘볼루션 디코더(CD)(320)보다 더 깊은 네트워크 구조를 가진 비대칭 이미지 압축 네트워크를 제안한다.

[0078] 콘볼루션 인코더(CE)(310)는 CNN(convolution neural network) 모델로서 콘볼루션 계층(Conv), GDN 계층, 스킵

연결(skip connection)을 가진 비대칭 콘볼루션 오토인코더 구조로 구성되며, 일례로 도 4에 도시한 바와 같이 콘볼루션 디코더(CD)(320)보다 더 깊은 네트워크 구성으로 첫 번째 콘볼루션 계층 이후 잔여 블록(residual block)(GB) 계층과 콘볼루션 계층의 결합이 여러 번 반복되어 구성된다. 예를 들어, 콘볼루션 디코더(CD)(320)의 네트워크 구성으로 디콘볼루션(deconvolution) 계층과 IGDN 계층의 결합이 2번 반복되는 경우 콘볼루션 인코더(CE)(310)는 잔여 블록(GB)과 콘볼루션 계층의 결합이 4번 반복되어 구성된다. 콘볼루션 인코더(CE)(310)는 이미지 압축을 위해 학습하고자 하는 매개변수에 따라 디코더(CD)(320)보다 더 깊은 네트워크 구조를 구성할 수 있다.

[0079] 그리고, 각 잔여 블록(GB)(511)은 도 5에 도시한 바와 같이 콘볼루션 계층(501), GDN 계층(502), 다른 콘볼루션 계층(503), 그리고 잔여 블록(GB)(511)의 입력과 출력 간의 스킵 연결(504)로 구성될 수 있으며, 이때 스킵 연결(504)은 입력 피처 맵을 일정 배수(예컨대, 2배수)로 축소한다.

[0080] 요컨대, 도 6에 도시한 바와 같이 기존에는 인코더의 네크워크 구조가 디코더와 서로 대칭되는 구조로 이루어지는 반면에, 본 발명에서의 콘볼루션 인코더(CE)(310)는 더 깊어진 네트워크 구조로 콘볼루션 계층과 잔여 블록(GB)이 콘볼루션 디코더(CD)(320)보다 더 많이 반복되면서 스킵 연결이 포함된 비대칭 구조로 이루어질 수 있다.

[0081] 또한, 본 발명에 따른 콘볼루션 인코더(CE)(310)에서 왜곡과 비율을 함께 고려함에 있어 이미지의 추가 정보를 활용하는 것 또한 가능하다. 일례로, 본 발명은 원본 이미지(301)의 추가 정보로서 원본 이미지(301)에서 분석된 주파수 정보(예컨대, DWT: discrete wavelet transform 등)가 콘볼루션 인코더(CE)(310)의 네트워크에 입력되는 구조를 포함할 수 있다. 예를 들어, 도 7에 도시한 바와 같이 콘볼루션 인코더(CE)(310)는 첫 번째 콘볼루션 계층에서 출력된 DWT 데이터가 차원 변경(예컨대, reshape 함수 등)을 거쳐 적어도 하나의 다음 콘볼루션 계층으로 입력되는 구조를 포함할 수 있다.

[0082] 콘볼루션 인코더(CE)(310)에서 이미지의 주파수 정보를 활용함으로써 압축 효율을 향상시킬 수 있고, 주파수 관점에서 서로 상이한 자연 이미지와 멀티미디어 이미지(예컨대, 웹툰 등)를 구분하여 효과적인 압축 성능을 실현할 수 있다.

[0083] 가중치 프루닝을 통한 경량 디코더

[0084] 가중치 프루닝은 콘볼루션 디코더(CD)(320)에 포함된 모든 콘볼루션 계층에 적용될 수 있다. 가중치 프루닝은 각 가중치의 크기가 프루닝 기준으로 설정될 수 있으며, 즉 크기가 작은 가중치를 덜 중요한 것으로 간주하여 프루닝을 수행할 수 있다.

[0085] 콘볼루션 디코더(CD)(320)에서는 다음의 단계를 거쳐 프루닝 과정을 수행할 수 있다. 먼저, 각 계층의 프루닝 민감도를 측정하여 계층 별 프루닝 임계값을 결정한다. 다음으로, 크기가 임계값보다 낮은 가중치를 제거한다. 마지막으로, 디코더 성능을 위해 프루닝된 네트워크를 재훈련한다. 이러한 프루닝 과정은 프루닝 효과를 극대화하기 위해 여러 번 반복될 수 있다.

[0086] 도 8은 본 발명의 일실시예에 있어서 가중치 프루닝 과정을 설명하기 위한 예시 도면이다.

[0087] 도 8을 참조하면, 이전 콘볼루션 계층의 출력(800)에 대해 해당 출력(800)을 입력으로 하는 다음 콘볼루션 계층의 각 채널(RGB 채널)에서 복수의 가중치를 가진 필터(81)를 통해 연산하게 되는데, 이때 가중치 프루닝은 필터가 가진 적어도 일부의 가중치를 제거하는 것을 목적으로 한다. 가중치 프루닝은 이전 콘볼루션 계층의 출력(800)을 입력으로 하는 다음 콘볼루션 계층에 대해 프루닝 임계값이 결정되면 필터(81)가 가진 가중치 중 크기가 프루닝 임계값 미만에 해당되는 가중치(8)를 제거한다.

[0088] 모든 필터(81)에 대해 프루닝 임계값 미만인 가중치(8)를 제거하는 가중치 프루닝을 수행함으로써 모델 사이즈 측면에서 보다 가벼운 콘볼루션 디코더(CD)(320)를 구축할 수 있어 메모리 성능을 향상시킬 수 있다.

[0089] 필터 프루닝을 통한 경량 디코더

[0090] 필터 프루닝은 필터가 가진 가중치 절대값의 평균 값의 크기가 프루닝 기준으로 설정될 수 있으며, 즉 가중치 절대값의 평균치가 큰 필터를 덜 중요한 것으로 간주하여 프루닝을 수행할 수 있다.

[0091] 필터 프루닝은 콘볼루션 디코더(CD)(320)에 포함된 콘볼루션 계층에 적용될 수 있다. 다만, 콘볼루션 디코더(CD)(320)의 마지막 계층은 세 개의 채널 즉, RGB 채널이 모두 있어야 재구성된 이미지 $\mathbf{I}_{Recon}$를 출력할 수 있

기 때문에 마지막 계층은 필터 프루닝 대상에서 제외될 수 있다.

[0092] 필터 프루닝을 위해서는 각 계층 별로 필터의 가중치 평균을 프루닝 기준으로 설정할 수 있으며, 각 계층의 프루닝 임계치와 제거해야 할 필터의 상대적인 양은 민감도 분석에 기초하여 실험을 통해 결정될 수 있다. 콘볼루션 디코더(CD)(320)에서 각 계층 별로 일부 필터가 제거된 후 디코더의 네트워크를 재훈련할 수 있다.

[0093] 도 9는 본 발명의 일실시예에 있어서 필터 프루닝 과정을 설명하기 위한 예시 도면이다.

[0094] 도 9를 참조하면, 이전 콘볼루션 계층의 출력(800)에 대해 해당 출력(800)을 입력으로 하는 다음 콘볼루션 계층의 각 채널(RGB 채널)에서 복수의 가중치를 가진 필터(81)를 통해 연산하게 되는데, 이때 필터 프루닝은 중요하지 않은 필터 전체를 제거하는 것을 목적으로 한다. 필터 프루닝은 이전 콘볼루션 계층의 출력(800)을 입력으로 하는 다음 콘볼루션 계층에 대해 프루닝 임계값이 결정되면 해당 계층의 필터(81) 중 가중치 평균치가 프루닝 임계값을 초과하는 필터(9)를 제거한다.

[0095] 가중치 평균이 임계값보다 작은 필터를 제거하는 것 또한 가능하나, 가중치 값이 작은 필터의 경우 이미지 분류를 위한 네트워크에서 기여도가 낮은 반면에, 이미지 압축을 위한 네트워크에서는 가중치 값이 큰 필터보다 더 중요한 특징을 가진다. 따라서, 본 발명에서의 필터 프루닝은 각 계층 별 필터(81) 중 가중치 평균이 프루닝 임계값보다 큰 필터(9)를 덜 중요한 것으로 간주하여 제거한다.

[0096] 마지막 계층을 제외한 모든 계층에 대해 프루닝 임계값보다 큰 필터(9)를 제거하는 필터 프루닝을 수행함으로써 디코딩 소요 시간 측면에서 고속의 콘볼루션 디코더(CD)(320)를 구축할 수 있어 계산 성능을 향상시킬 수 있다.

[0097] 이처럼 본 발명의 실시예들에 따르면, 본 발명의 실시예들에 따르면, 딥러닝 기반 이미지 압축 기술에서 디코더에 비해 더 깊은 네트워크로 설계된 인코더의 비대칭 구조를 통해 이미지의 비율 왜곡 성능을 향상시킬 수 있다. 더 나아가, 본 발명의 실시예들에 따르면, 이미지 압축 기술에서 디코더에 프루닝 기법을 적용함으로써 모델 사이즈 측면에서 더욱 가볍고 처리 시간과 속도 측면에서 보다 빠른 디코더를 제공할 수 있다.

[0098] 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, 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, 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 어플리케이션을 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.

[0099] 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, 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.

[0100] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 이때, 매체는 컴퓨터로 실행 가능한 프로그램을 계속 저장하거나, 실행 또는 다운로드를 위해 임시 저장하는 것일 수도 있다. 또한, 매체는 단일 또는 수 개의 하드웨어가 결합된 형태의 다양한 기록수단 또는 저장수단일 수 있는데, 어떤 컴퓨터 시스템에 직접 접속되는 매체에 한정되지 않고, 네트워크 상에 분산 존재하는 것일 수도 있다. 매체의 예시로는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM 및 DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical medium), 및 ROM, RAM, 플래시 메모리 등을 포함하여 프로그램 명령어가 저장되도록 구성된 것이 있을 수 있다. 또한, 다른 매체의 예시로, 어플리케이션을 유통하는 앱 스토어나 기타 다양한 소프트웨어

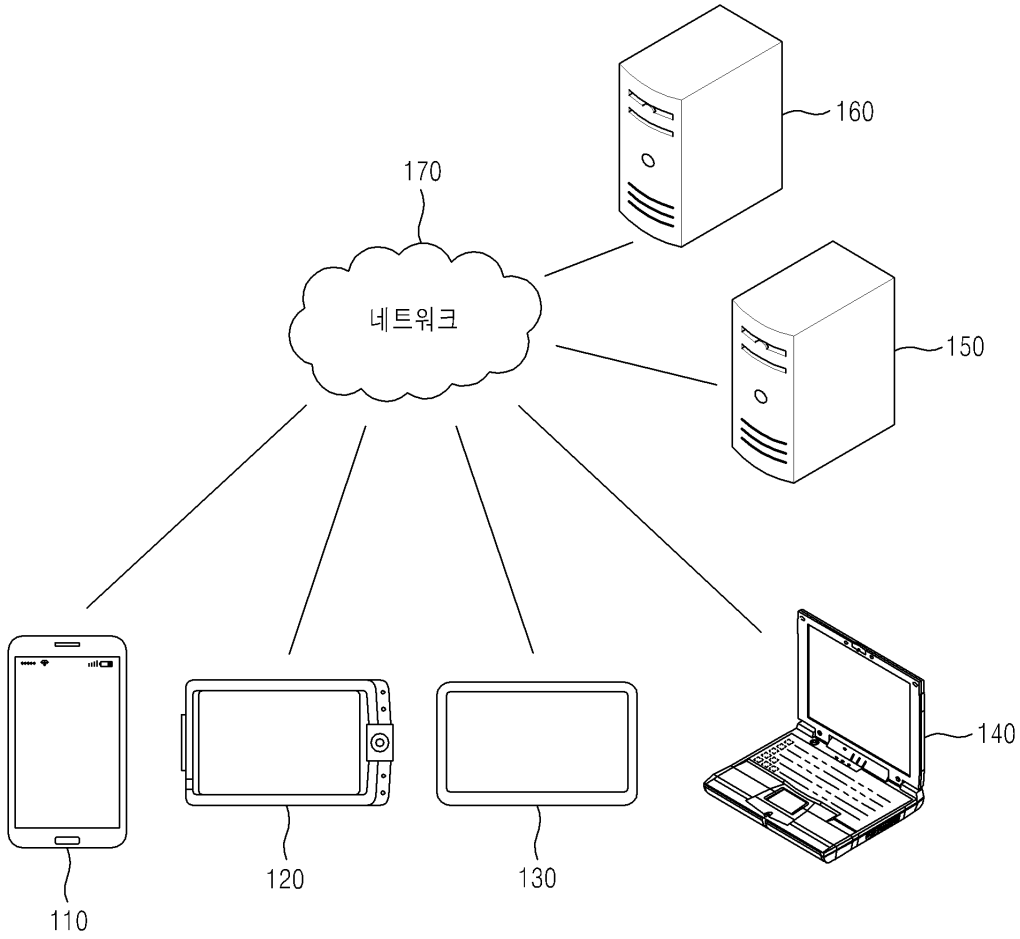
를 공급 내지 유통하는 사이트, 서버 등에서 관리하는 기록매체 내지 저장매체도 들 수 있다.

[0101] 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.

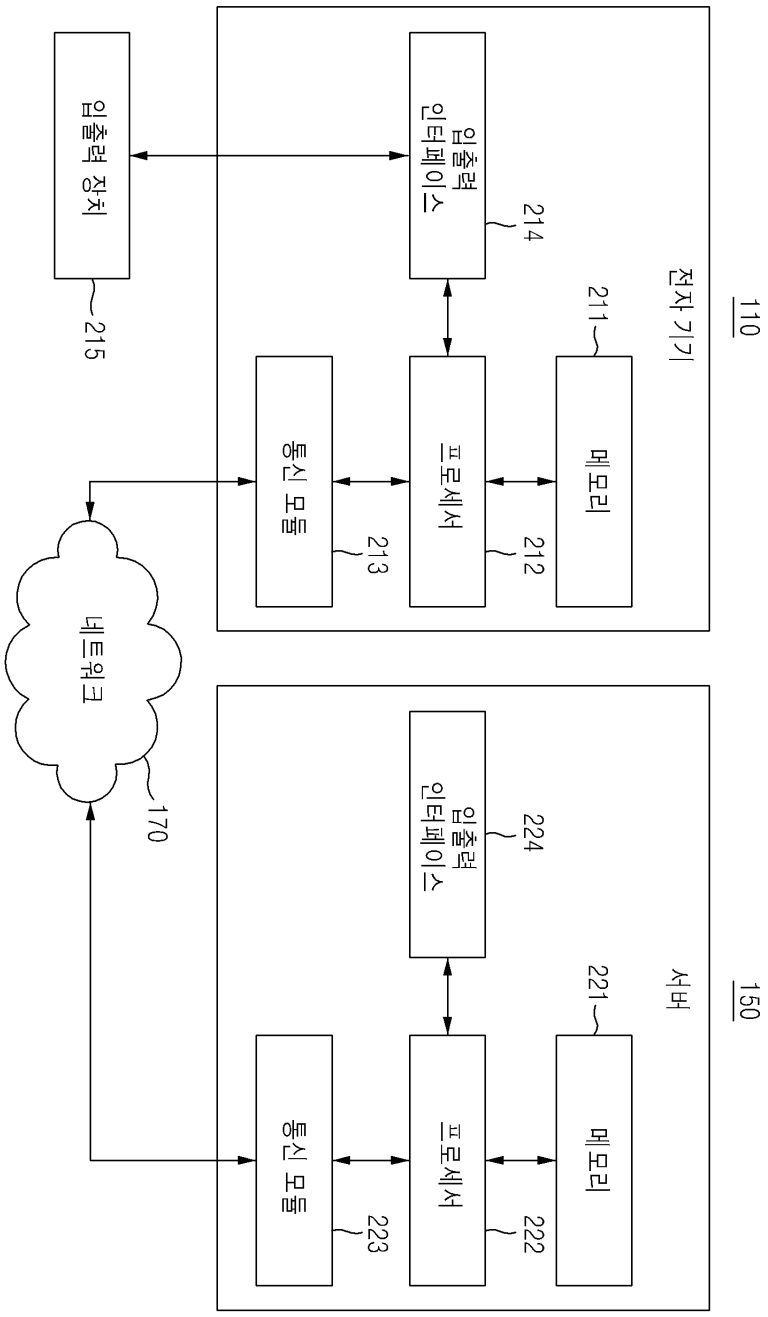
[0102] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

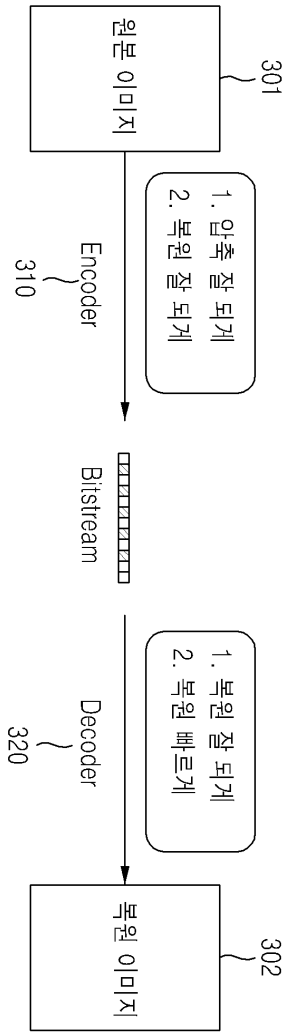
도면1



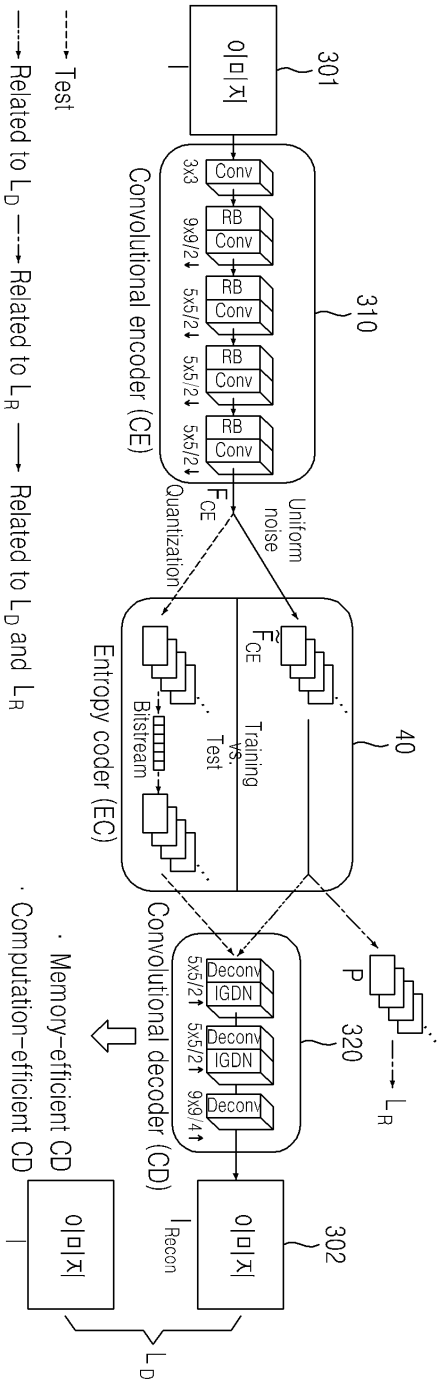
도면2


입출력 장치
215
입출력 인터페이스
214
전자 기기
110
211
메모리
프로세서
212
통신 모듈
213
네트워크
170
입출력 인터페이스
224
서버
150
221
메모리
프로세서
222
통신 모듈
223

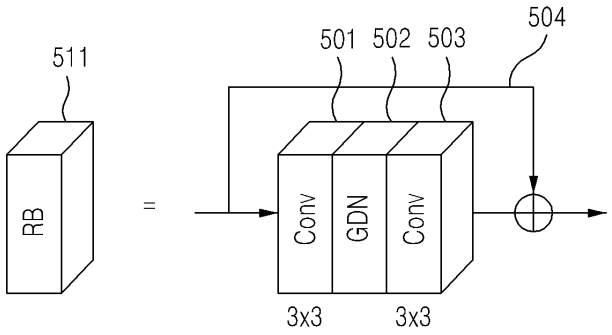
도면3


원본 이미지
301
1. 압축 잘 되게
2. 복원 잘 되게
Encoder
310
Bitstream
1. 복원 잘 되게
2. 복원 빠르게
Decoder
320
복원 이미지
302

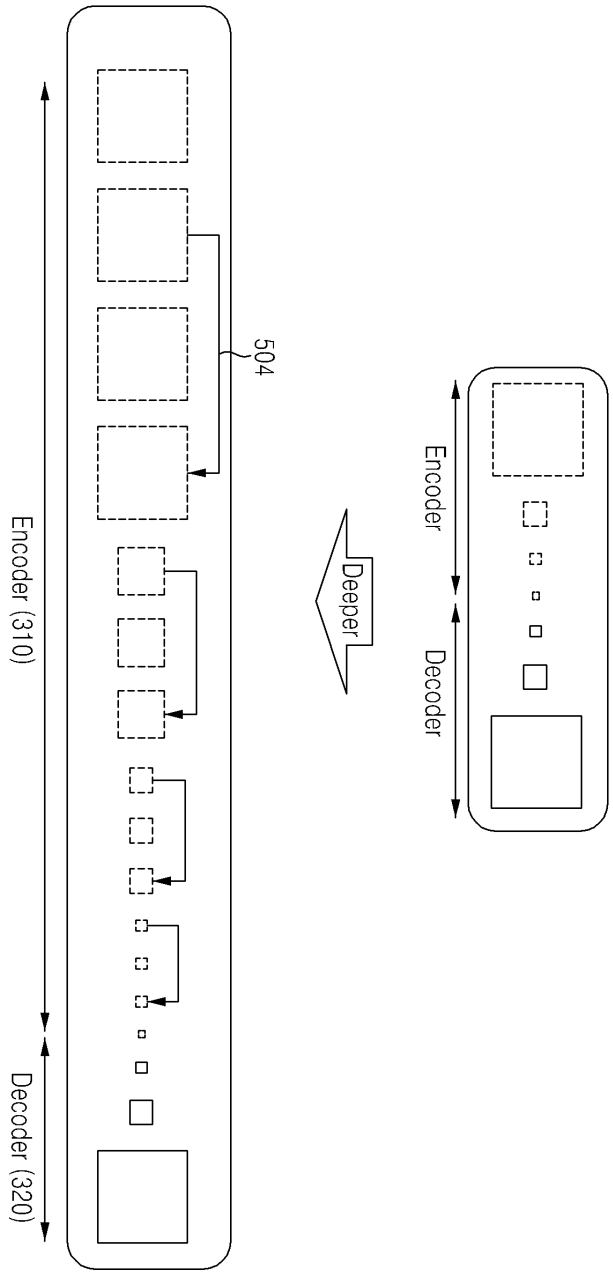
도면4


이미지
301
Convolutional encoder (CE)
3x3
Conv
9x9/2↓
RB
Conv
5x5/2↓
RB
Conv
5x5/2↓
RB
Conv
5x5/2↓
RB
Conv
310
Quantization
Uniform noise
Entropy coder (EC)
Bitstream
Test
Training vs. Test
40
Convolutional decoder (CD)
5x5/2↑
Deconv
IGDN
5x5/2↑
Deconv
IGDN
9x9/4↑
Deconv
320
P
이미지
302
Test
Related to L_D
Related to L_R
Related to L_D and L_R
· Memory-efficient CD
· Computation-efficient CD

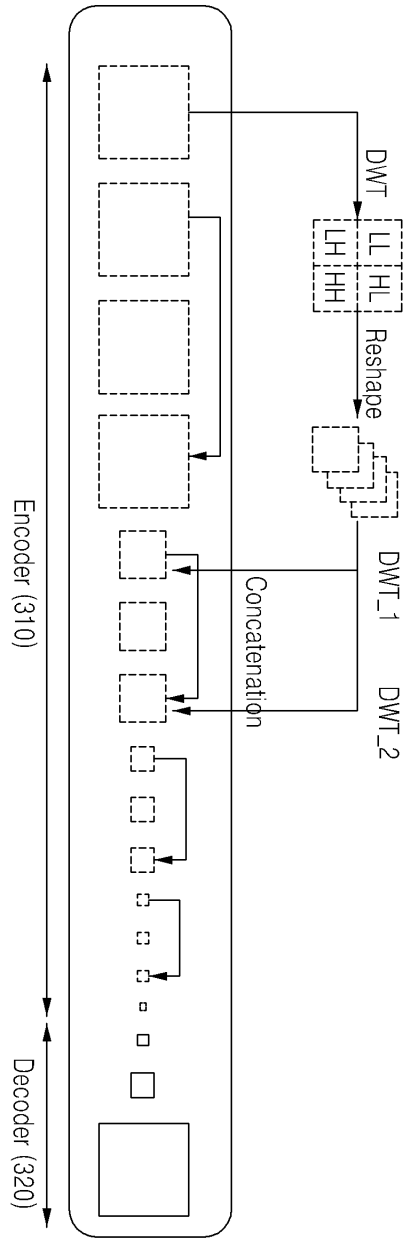
도면5


511
RB
=
501
502
503
504
Conv
GDN
Conv
3x3
3x3

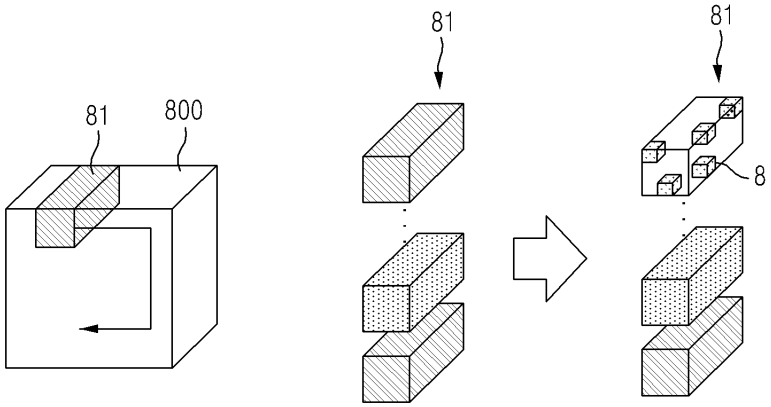
도면6


Encoder (310)
Decoder (320)
504
Deeper
Encoder
Decoder

도면7


Encoder (310)
Decoder (320)
DWT
LL
HL
LH
HH
Reshape
DWT_1
DWT_2
Concatenation

도면8



도면9