(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0072450
(43) 공개일자 2019년06월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2018-0160774
(22) 출원일자 2018년12월13일
심사청구일자 없음
(30) 우선권주장
1020170172827 2017년12월15일 대한민국(KR)

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
조승현
대전광역시 유성구 배울2로 61, 1014동 103호
김연희
대전광역시 유성구 가정로 270, 1동 323호
(뒷면에 계속)
(74) 대리인
한양특허법인

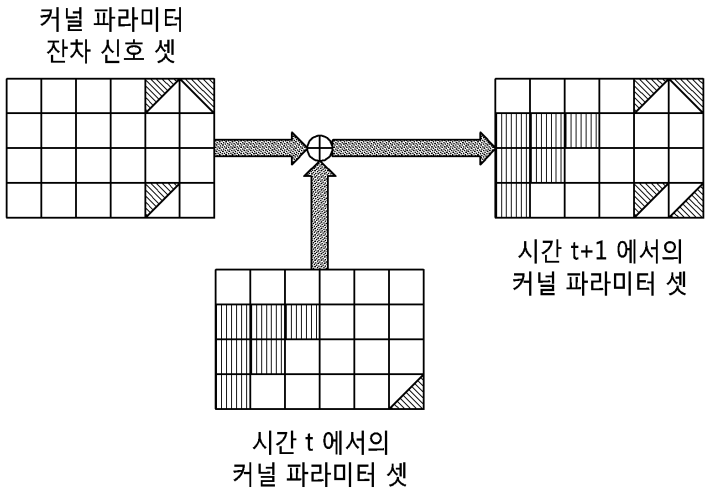
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **분산 처리 환경에서의 학습 파라미터의 압축 및 전송을 제공하는 방법 및 장치**

(57) 요 약

분산 처리 환경에서의 딥 러닝 모델의 학습을 위한 학습 파라미터를 압축하고, 압축된 파라미터를 전송하는 방법 및 장치가 개시된다. 분산 처리 시스템 내의 복수의 전자 장치들은 신경망의 학습을 수행한다. 학습의 수행에 의해 파라미터가 갱신된다. 전자 장치는 자신의 갱신된 파라미터를 다른 전자 장치들과 공유할 수 있다. 이러한 공유를 효율적으로 수행하기 위해, 파라미터의 잔차가 다른 전자 장치에게 제공된다. 파라미터의 잔차가 제공되면, 다른 전자 장치는 파라미터의 잔차를 사용하여 파라미터의 갱신을 수행한다.

대 표 도 - 도8



(72) 발명자

석진욱
대전광역시 유성구 배울1로 13, 201동 1002호

이주영
대전광역시 유성구 문화원로 13, 107동 804호

임웅
대전광역시 유성구 노은서로266번길 61, 603호

김종호
대전광역시 유성구 지족북로 60, 205동 2001호

이대열
대전광역시 유성구 신성남로 69-18

정세윤
대전광역시 유성구 지족북로 60, 207동 1401호

김휘용
대전광역시 유성구 은구비남로 34, 810동 201호

최진수
대전광역시 유성구 반석서로 98, 609동 1605호

강제원
서울특별시 서대문구 이화여대길 52, 아산공학관 432호

이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-00072
부처명	과학기술정보통신부
연구관리전문기관	정보통신기술진흥센터(IITP)
연구사업명	ETRI연구개발지원사업
연구과제명	초실감 테라미디어를 위한 AV부호화 및 LF미디어 원천기술 개발
기 여 율	1/1
주관기관	한국전자통신연구원
연구기간	2017.01.01 ~ 2017.12.31

명 세 서

청구범위

청구항 1

분산 처리 시스템 내의 전자 장치에 의해 수행되는,

신경망의 파라미터의 갱신을 위한 정보를 생성하는 단계; 및

상기 파라미터의 갱신을 위한 정보를 전송하는 단계

를 포함하고,

상기 파라미터의 갱신을 위한 정보는,

학습의 수행에 의해 갱신된 상기 파라미터의 상태 및 상기 학습의 이전의 상기 파라미터의 상태에 기반하여 생성되는, 갱신된 파라미터에 대한 정보를 제공하는 방법.

청구항 2

제1항에 있어서,

상가 파라미터의 갱신을 위한 정보는 상기 파라미터의 잔차에 기반하여 생성되고,

상기 파라미터의 잔차는 상기 학습의 수행에 의해 갱신된 상기 파라미터의 값 및 상기 학습의 이전의 상기 파라미터의 값 간의 차이인, 갱신된 파라미터에 대한 정보를 제공하는 방법.

청구항 3

제1항에 있어서,

상기 파라미터의 갱신을 위한 정보는 상기 분산 처리 시스템 내에서 상기 신경망의 학습을 수행하는 하나 이상의 다른 전자 장치로 전송되는, 갱신된 파라미터에 대한 정보를 제공하는 방법.

청구항 4

제1항에 있어서,

상기 파라미터의 갱신을 위한 정보는 상기 분산 처리 시스템 내의 파라미터 서버로 전송되고,

상기 파라미터 서버는 상기 파라미터의 갱신을 위한 정보를 상기 분산 처리 시스템 내에서 상기 신경망의 학습을 수행하는 하나 이상의 전자 장치들로 전달하는, 갱신된 파라미터에 대한 정보를 제공하는 방법.

청구항 5

제1항에 있어서,

상기 전자 장치는 딥 러닝 모델의 복수의 레이어들 중 하나의 레이어에 대한 학습을 수행하는, 갱신된 파라미터에 대한 정보를 제공하는 방법.

청구항 6

분산 처리 시스템 내의 전자 장치에 의해 수행되는,

상기 분산 처리 시스템 내의 다른 장치로부터 파라미터의 갱신을 위한 정보를 수신하는 단계; 및

상기 파라미터의 갱신을 위한 정보를 사용하여 상기 파라미터의 갱신을 수행하는 단계

를 포함하고,

상기 파라미터의 갱신을 위한 정보에 기반하여 파라미터의 잔차가 생성되고,

상기 파라미터의 잔차를 상기 파라미터에 더함으로써 상기 파라미터가 갱신되는, 파라미터 갱신 방법.

청구항 7

제6항에 있어서,

상기 파라미터는 복수인, 파라미터 갱신 방법.

청구항 8

제7항에 있어서,

상기 복수의 파라미터들은 딥 러닝 모델의 하나의 계층을 구성하는 딥 러닝 파라미터들인, 파라미터 갱신 방법.

청구항 9

제7항에 있어서,

상기 파라미터의 갱신을 위한 정보는 상기 복수의 파라미터들의 압축된 잔차들을 포함하는, 파라미터 갱신 방법.

청구항 10

제7항에 있어서,

파라미터의 갱신을 위한 정보에 포함된 상기 복수의 파라미터들의 압축된 잔차들에 대한 복호화를 수행함으로써 상기 복수의 파라미터들의 잔차들이 획득되고, 상기 복수의 파라미터들의 잔차들을 사용하여 상기 복수의 파라미터들의 갱신이 수행되는, 파라미터 갱신 방법.

청구항 11

제10항에 있어서,

상기 복호화에 있어서, 영상의 블록에 대한 복호화 방법이 사용되는, 파라미터 갱신 방법.

청구항 12

제11항에 있어서,

상기 복수의 파라미터들의 각 파라미터는 상기 블록에 대응하는, 파라미터 생성 방법.

청구항 13

제11항에 있어서,

상기 블록은 복수이고,

상기 복수의 블록들은 대상 영상을 구성하고,

상기 복호화에 있어서, 대상 영상에 대한 복호화 방법이 사용되는, 파라미터 갱신 방법.

청구항 14

제11항에 있어서,

상기 복호화를 위해 상기 블록에 대한 엔트로피 복호화, 스캐닝, 역양자화 및 역변환 중 하나 이상이 사용되는, 파라미터 갱신 방법.

청구항 15

제7항에 있어서,

상기 파라미터의 갱신을 위한 정보에 기반하여 스캐닝된 정보가 생성되고,

상기 스캐닝된 정보는 스캐닝된 양자화된 경사들을 포함하는, 파라미터 갱신 방법.

청구항 16

제15항에 있어서,

상기 스캐닝된 정보에 대한 1차원으로부터 2차원으로의 스캐닝을 수행함으로써 양자화된 경사들이 생성되고,

상기 양자화된 경사들에 대한 역양자화를 수행함으로써 재구축된 경사들이 생성되고,

상기 재구축된 경사들은 상기 복수의 파라미터들의 잔차들인, 파라미터 갱신 방법.

청구항 17

제15항에 있어서,

상기 스캐닝된 정보는 제로 서브 블록을 포함하고,

상기 제로 서브 블록은 블록 내의 특정된 영역을 나타내고, 상기 특정된 영역 내의 양자화된 경사들의 값들이 모두 0임을 나타내는, 파라미터 갱신 방법.

청구항 18

제15항에 있어서,

상기 스캐닝된 정보는 제로 블록을 포함하고,

상기 제로 블록은 블록 내의 양자화된 경사들의 값들이 모두 0임을 나타내는, 파라미터 갱신 방법.

청구항 19

제15항에 있어서,

상기 스캐닝된 정보는 제로 레이어를 포함하고,

상기 제로 레이어는 레이어의 모든 블록들의 모든 양자화된 경사들의 값들이 0임을 나타내는, 파라미터 갱신 방법.

청구항 20

분산 처리 시스템 내의 전자 장치에 있어서,

상기 분산 처리 시스템 내의 다른 장치로부터 파라미터의 갱신을 위한 정보를 수신하는 통신부; 및

상기 파라미터의 갱신을 위한 정보를 사용하여 상기 파라미터의 갱신을 수행하는 처리부

를 포함하고,

상기 처리부는 파라미터의 갱신을 위한 정보에 기반하여 파라미터의 잔차를 생성하고, 상기 파라미터의 잔차를 상기 파라미터에 더함으로써 상기 파라미터의 갱신을 수행하는, 전자 장치.

발명의 설명

기 술 분 야

[0001] 아래의 실시예들은 분산 처리 환경에서의 딥 러닝을 위한 방법 및 장치에 관한 것으로서, 분산 처리 환경에서의 딥 러닝 모델의 학습을 위한 학습 파라미터를 압축하고, 압축된 파라미터를 전송하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 컨벌루션 신경망(Convolutional Neural Network; CNN)과 같은 최첨단의(state-of-the-art) 딥 러닝 모델의 특징은 인식 성능을 향상시키기 위해 대규모의 네트워크 구조의 형태를 갖는 것이다.

[0003] 대규모의 네트워크 구조에서는, 모델의 계층(hierarchical) 깊이가 점점 더 증가한다. 깊이의 증가에 따라 네트워크의 노드들을 연결하는 에지의 가중치(weight) 파라미터들의 개수 또한 증가한다.

[0004] 대규모의 딥 러닝 모델을 개발하기 위해서는, 분산 처리에 기반하는 고성능 컴퓨터 인프라가 반드시 필요하다. 말하자면, 입력 데이터의 양이 많아질수록 수 많은 컴퓨터들이 필요하고, 수 많은 컴퓨터들이 사용됨에 따라 효율적인 분산 학습 알고리즘이 중요하다.

[0005] 분산 학습의 가속화 방식은 데이터 병렬처리(data parallelism) 방식 및 모델 병렬처리(model parallelism) 방식으로 구분될 수 있다.

[0006] 데이터 병렬처리 방식은 학습의 대상인 입력 데이터 셋(set)이 다수의 컴퓨터들에게 분할되고, 각 컴퓨터가 자신에게 할당된 분할된 입력 데이터 셋을 사용하는 학습을 수행하는 방식이다.

[0007] 모델 병렬처리 방식은 딥 러닝 모델이 분할되고, 다수의 컴퓨터가 분할된 딥 러닝 모델들의 대한 학습을 각각 수행하는 방식이다.

[0008] 보다 구체적으로, 데이터 병렬처리 방식에서, 분산 처리 시스템의 각 분산 컴퓨터(distributed computer)는 부분으로 나뉜 입력 데이터를 사용하여 딥 러닝 학습 모델의 전체에 대한 학습을 수행한다.

[0009] 이 때, 학습의 반복(iteration)이 수행될 때마다, 분산 컴퓨터는 학습에 의해 수정된 커널 파라미터를 다른 분산 컴퓨터와 교환한다. 말하자면, 분산 처리 시스템의 분산 컴퓨터들 간에서 수정된 커널 파라미터들의 교환이 이루어진다.

[0010] 한편, 대규모의 딥 러닝 모델이 하나의 컴퓨터에 의해 로드 및 처리될 수 없을 정도가 되면, 딥 러닝 모델을 분할하고, 분할된 딥 러닝 모델에 대한 학습을 수행하기 위해 데이터 병렬처리 방식이 사용될 수 있다. 데이터 병렬처리 방식에서, 분산 처리 시스템의 다수의 분산 컴퓨터들은 딥 러닝 모델의 분할들에 대한 학습을 각각 수행한다.

[0011] 이 때, 분산 처리 시스템의 각 분산 컴퓨터에서는 전체의 입력 데이터에 대해 트레이닝을 수행하고, 딥 러닝 모델의 분할에 대해 부분적으로 계산된 로컬 파라미터를 다른 분산 컴퓨터와 교환한다. 말하자면, 분산 처리 시스템의 분산 컴퓨터들 간에서 로컬 파라미터의 교환이 이루어진다.

[0012] 즉. 분산 처리에 기반하는 딥 러닝 모델의 학습 과정에 있어서, 분산 처리 시스템의 분산 컴퓨터들이 커널 파라미터를 서로 교환하는 과정이 필수적으로 요구된다.

발명의 내용

해결하려는 과제

[0013] 일 실시예는 분산 처리 환경에서의 학습 파라미터의 압축 및 전송을 제공하는 방법 및 장치를 제공할 수 있다.

[0014] 일 실시예는 하나의 전자 장치에서 갱신된 파라미터를 분산 처리 시스템 내의 복수의 전자 장치들 간에 공유하는 방법 및 장치를 제공할 수 있다.

과제의 해결 수단

[0015] 일 측에 있어서, 분산 처리 시스템 내의 전자 장치에 의해 수행되는, 신경망의 파라미터의 갱신을 위한 정보를 생성하는 단계; 및 상기 파라미터의 갱신을 위한 정보를 전송하는 단계를 포함하고, 상기 파라미터의 갱신을 위한 정보는, 학습의 수행에 의해 갱신된 상기 파라미터의 상태 및 상기 학습의 이전의 상기 파라미터의 상태에 기반하여 생성되는, 갱신된 파라미터에 대한 정보를 제공하는 방법이 제공된다.

[0016] 상가 파라미터의 갱신을 위한 정보는 상기 파라미터의 잔차에 기반하여 생성될 수 있다.

[0017] 상기 파라미터의 잔차는 상기 학습의 수행에 의해 갱신된 상기 파라미터의 값 및 상기 학습의 이전의 상기 파라미터의 값 간의 차이일 수 있다.

[0018] 상기 파라미터의 갱신을 위한 정보는 상기 분산 처리 시스템 내에서 상기 신경망의 학습을 수행하는 하나 이상의 다른 전자 장치로 전송될 수 있다.

[0019] 상기 파라미터의 갱신을 위한 정보는 상기 분산 처리 시스템 내의 파라미터 서버로 전송될 수 있다.

[0020] 상기 파라미터 서버는 상기 파라미터의 갱신을 위한 정보를 상기 분산 처리 시스템 내에서 상기 신경망의 학습을 수행하는 하나 이상의 전자 장치들로 전달할 수 있다.

[0021] 상기 전자 장치는 딥 러닝 모델의 복수의 레이어들 중 하나의 레이어에 대한 학습을 수행할 수 있다.

[0022] 다른 일 측에 있어서, 분산 처리 시스템 내의 전자 장치에 의해 수행되는, 상기 분산 처리 시스템 내의 다른 장치로부터 파라미터의 갱신을 위한 정보를 수신하는 단계; 및 상기 파라미터의 갱신을 위한 정보를 사용하여 상기 파라미터의 갱신을 수행하는 단계를 포함하고, 상기 파라미터의 갱신을 위한 정보에 기반하여 파라미터의 잔차가 생성되고, 상기 파라미터의 잔차를 상기 파라미터에 더함으로써 상기 파라미터가 갱신되는, 파라미터 갱신 방법이 제공된다.

[0023] 상기 파라미터는 복수일 수 있다.

[0024] 상기 복수의 파라미터들은 딥 러닝 모델의 하나의 계층을 구성할 수 있다.

[0025] 상기 파라미터의 갱신을 위한 정보는 상기 복수의 파라미터들의 압축된 잔차들을 포함할 수 있다.

[0026] 파라미터의 갱신을 위한 정보에 포함된 상기 복수의 파라미터들의 압축된 잔차들에 대한 복호화를 수행함으로써 상기 복수의 파라미터들의 잔차들이 획득될 수 있다.

[0027] 상기 복수의 파라미터들의 잔차들을 사용하여 상기 복수의 파라미터들의 갱신이 수행될 수 있다.

[0028] 상기 복호화에 있어서, 영상의 블록에 대한 복호화 방법이 사용될 수 있다.

[0029] 상기 복수의 파라미터들의 각 파라미터는 상기 블록에 대응할 수 있다.

[0030] 상기 블록은 복수일 수 있다.

[0031] 상기 복수의 블록들은 대상 영상을 구성할 수 있다.

[0032] 상기 복호화에 있어서, 대상 영상에 대한 복호화 방법이 사용될 수 있다.

[0033] 상기 복호화를 위해 상기 블록에 대한 엔트로피 복호화, 스캐닝, 역양자화 및 역변환 중 하나 이상이 사용될 수 있다.

[0034] 상기 파라미터의 갱신을 위한 정보에 기반하여 스캐닝된 정보가 생성될 수 있다.

[0035] 상기 스캐닝된 정보는 스캐닝된 양자화된 경사들을 포함할 수 있다.

[0036] 상기 스캐닝된 정보에 대한 1차원으로부터 2차원으로의 스캐닝을 수행함으로써 양자화된 경사들이 생성될 수 있다.

[0037] 상기 양자화된 경사들에 대한 역양자화를 수행함으로써 재구축된 경사들이 생성될 수 있다.

[0038] 상기 재구축된 경사들은 상기 복수의 파라미터들의 잔차들일 수 있다.

[0039] 상기 스캐닝된 정보는 제로 서브 블록을 포함할 수 있다.

[0040] 상기 제로 서브 블록은 블록 내의 특정된 영역을 나타낼 수 있고, 상기 특정된 영역 내의 양자화된 경사들의 값들이 모두 0임을 나타낼 수 있다.

[0041] 상기 스캐닝된 정보는 제로 블록을 포함할 수 있다.

[0042] 상기 제로 블록은 블록 내의 양자화된 경사들의 값들이 모두 0임을 나타낼 수 있다.

[0043] 상기 스캐닝된 정보는 제로 레이어를 포함할 수 있다.

[0044] 상기 제로 레이어는 레이어의 모든 블록들의 모든 양자화된 경사들의 값들이 0임을 나타낼 수 있다.

[0045] 또 다른 일 측에 있어서, 분산 처리 시스템 내의 전자 장치에 있어서, 상기 분산 처리 시스템 내의 다른 장치로부터 파라미터의 갱신을 위한 정보를 수신하는 통신부; 및 상기 파라미터의 갱신을 위한 정보를 사용하여 상기 파라미터의 갱신을 수행하는 처리부를 포함하고, 상기 처리부는 파라미터의 갱신을 위한 정보에 기반하여 파라미터의 잔차를 생성하고, 상기 파라미터의 잔차를 상기 파라미터에 더함으로써 상기 파라미터의 갱신을 수행하는, 전자 장치가 제공된다.

발명의 효과

[0046] 분산 처리 환경에서의 학습 파라미터의 압축 및 전송을 제공하는 방법 및 장치가 제공된다.

[0047] 하나의 전자 장치에서 갱신된 파라미터를 분산 처리 시스템 내의 복수의 전자 장치들 간에 공유하는 방법 및 장치가 제공된다.

도면의 간단한 설명

[0048] 도 1은 본 발명이 적용되는 부호화 장치의 일 실시예에 따른 구성을 나타내는 블록도이다.

도 2는 본 발명이 적용되는 복호화 장치의 일 실시예에 따른 구성을 나타내는 블록도이다.

도 3은 영상을 부호화 및 복호화할 때의 영상의 분할 구조를 개략적으로 나타내는 도면이다.

도 4는 일 예에 따른 풀 메시 토폴로지에 기반하는 파라미터의 공유를 나타낸다.

도 5는 일 예에 따른 스타 토폴로지에 기반하는 파라미터의 공유를 나타낸다.

도 6은 일 실시예에 따른 갱신된 파라미터에 대한 정보를 제공하는 방법의 흐름도이다.

도 7은 일 실시예에 따른 갱신된 파라미터에 대한 정보를 사용하여 파라미터의 갱신을 수행하는 방법의 흐름도이다.

도 8은 일 예에 따른 딥 러닝의 파라미터 셋의 갱신을 나타낸다.

도 9는 일 예에 따른 양자화, 레스터 스캔 및 엔트로피 코딩의 과정을 나타낸다.

도 10은 일 예에 따른 제로 서브 블록 및 제로 레이어를 사용하는 엔트로피 부호화를 나타낸다.

도 11은 일 예에 따른 모델 병렬 처리 방식에서의 파라미터의 공유를 나타낸다.

도 12는 일 예에 따른 데이터 병렬 처리 방식에서의 파라미터의 공유를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0049] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 그러나, 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0050] 후술하는 예시적 실시예들에 대한 상세한 설명은, 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 실시예를 실시할 수 있기에 충분하도록 상세히 설명된다. 다양한 실시예들은 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들면, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예에 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 실시예의 정신 및 범위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 예시적 실시예들의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다.

[0051] 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다. 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0052] 본 발명에서 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들면, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

[0053] 어떤 구성요소(component)가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기의 2개의 구성요소들이 서로 간에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있으나, 상기의 2개의 구성요소들의 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 어떤 구성요소(component)가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 상기의 2개의 구성요소들의 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0054] 본 발명의 실시예에 나타나는 구성요소들은 서로 다른 특징적인 기능들을 나타내기 위해 독립적으로 도시되는 것으로, 각 구성요소들이 분리된 하드웨어나 하나의 소프트웨어 구성단위로 이루어짐을 의미하지 않는다. 즉, 각 구성요소는 설명의 편의상 각각의 구성요소로 나열하여 포함한 것으로 각 구성요소 중 적어도 두 개의 구성요소가 합쳐져 하나의 구성요소로 이루어지거나, 하나의 구성요소가 복수 개의 구성요소로 나뉘어져 기능을 수행할 수 있고 이러한 각 구성요소의 통합된 실시예 및 분리된 실시예도 본 발명의 본질에서 벗어나지 않는 한 본 발명의 권리범위에 포함된다.

[0055] 또한, 예시적 실시예들에서 특정 구성을 "포함"한다고 기술하는 내용은 상기의 특정 구성 이외의 구성을 배제하는 것이 아니며, 추가적인 구성이 예시적 실시예들의 실시 또는 예시적 실시예들의 기술적 사상의 범위에 포함될 수 있음을 의미한다.

[0056] 본 발명에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 발명에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다. 즉, 본 발명에서 특정 구성을 "포함"한다고 기술하는 내용은 해당 구성 이외의 구성을 배제하는 것이 아니며, 추가적인 구성 또한 본 발명의 실시 또는 본 발명의 기술적 사상의 범위에 포함될 수 있음을 의미한다.

[0057] 이하에서는, 기술분야에서 통상의 지식을 가진 자가 실시예들을 용이하게 실시할 수 있도록 하기 위하여, 첨부된 도면을 참조하여 실시 형태에 대하여 구체적으로 설명한다. 실시예들을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 명세서의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다. 또한, 도면 상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고, 동일한 구성요소에 대한 중복된 설명은 생략한다.

[0058] 이하에서, 영상은 비디오(video)을 구성하는 하나의 픽처(picture)를 의미할 수 있으며, 비디오 자체를 나타낼 수도 있다. 예를 들면, "영상의 부호화 및/또는 복호화"는 "비디오의 부호화 및/또는 복호화"를 의미할 수 있으며, "비디오를 구성하는 영상들 중 하나의 영상의 부호화 및/또는 복호화"를 의미할 수도 있다.

[0059] 이하에서, 용어들 "비디오(video)" 및 "동영상(motion picture)"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0060] 이하에서, 대상 영상은 부호화의 대상인 부호화 대상 영상 및/또는 복호화의 대상인 복호화 대상 영상일 수 있다. 또한, 대상 영상은 부호화 장치로 입력된 입력 영상일 수 있고, 복호화 장치로 입력된 입력 영상일 수 있다.

[0061] 이하에서, 용어들 "영상", "픽처", "프레임(frame)" 및 "스크린(screen)"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0062] 이하에서, 대상 블록은 부호화의 대상인 부호화 대상 블록 및/또는 복호화의 대상인 복호화 대상 블록일 수 있다. 또한, 대상 블록은 현재 부호화 및/또는 복호화의 대상인 현재 블록일 수 있다. 예를 들면, 용어들 "대상 블록" 및 "현재 블록"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0063] 이하에서, 용어들 "블록" 및 "유닛"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다. 또는 "블록"은 특정한 유닛을 나타낼 수 있다.

[0064] 이하에서, 용어들 "영역(region)" 및 "세그먼트(segment)"는 서로 교체되어 사용될 수 있다.

[0065] 이하에서, 특정한 신호는 특정한 블록을 나타내는 신호일 수 있다. 예를 들면, 원(original) 신호는 대상 블록을 나타내는 신호일 수 있다. 예측(prediction) 신호는 예측 블록을 나타내는 신호일 수 있다. 잔차(residual) 신호는 잔차 블록을 나타내는 신호일 수 있다.

[0066] 실시예들에서, 특정된 정보, 데이터, 플래그(flag) 및 요소(element), 속성(attribute) 등의 각각은 값을 가질 수 있다. 정보, 데이터, 플래그(flag) 및 요소(element), 속성(attribute) 등의 값 "0"은 논리 거짓(logical false) 또는 제1 기정의된(predefined) 값을 나타낼 수 있다. 말하자면, 값 "0", 거짓, 논리 거짓 및 제1 기정의된 값은 서로 대체되어 사용될 수 있다. 정보, 데이터, 플래그(flag) 및 요소(element), 속성(attribute) 등의 값 "1"은 논리 참(logical true) 또는 제2 기정의된(predefined) 값을 나타낼 수 있다. 말하자면, 값 "1",

참, 논리 참 및 제2 기정의된 값은 서로 대체되어 사용될 수 있다.

[0067] 행, 열 또는 인덱스(index)를 나타내기 위해 i 또는 j 등의 변수가 사용될 때, i의 값은 0 이상의 정수일 수 있으며, 1 이상의 정수일 수도 있다. 말하자면, 실시예들에서 행, 열 및 인덱스 등은 0에서부터 카운트될 수 있으며, 1에서부터 카운트될 수 있다.

[0068] 아래에서는, 실시예들에서 사용되는 용어가 설명된다.

[0069] 부호화기(encoder): 부호화(encoding)를 수행하는 장치를 의미한다.

[0070] 복호화기(decoder): 복호화(decoding)를 수행하는 장치를 의미한다.

[0071] 유닛(unit): 유닛은 영상의 부호화 및 복호화의 단위를 나타낼 수 있다. 용어들 "유닛" 및 "블록(block)"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0072] - 유닛은 샘플의 MxN 배열일 수 있다. M 및 N은 각각 양의 정수일 수 있다. 유닛은 흔히 2차원의 샘플들의 배열을 의미할 수 있다.

[0073] - 영상의 부호화 및 복호화에 있어서, 유닛은 하나의 영상의 분할에 의해 생성된 영역일 수 있다. 말하자면, 유닛은 하나의 영상 내의 특정된 영역일 수 있다. 하나의 영상은 복수의 유닛들로 분할될 수 있다. 또는, 유닛은 하나의 영상을 세분화된 부분들로 분할하고, 분할된 부분에 대한 부호화 또는 복호화가 수행될 때, 상기의 분할된 부분을 의미할 수 있다.

[0074] - 영상의 부호화 및 복호화에 있어서, 유닛의 종류에 따라서 유닛에 대한 기정의된 처리가 수행될 수 있다.

[0075] - 기능에 따라서, 유닛의 타입은 매크로 유닛(Macro Unit), 코딩 유닛(Coding Unit; CU), 예측 유닛(Prediction Unit; PU), 잔차 유닛(Residual Unit) 및 변환 유닛(Transform Unit; TU) 등으로 분류될 수 있다. 또는, 기능에 따라서, 유닛은 블록, 매크로블록(Macroblock), 코딩 트리 유닛(Coding Tree Unit), 코딩 트리 블록(Coding Tree Block), 코딩 유닛(Coding Unit), 부호화 블록(Coding Block), 예측 유닛(Prediction Unit), 예측 블록(Prediction Block), 잔차 유닛(Residual Unit), 잔차 블록(Residual Block), 변환 유닛(Transform Unit) 및 변환 블록(Transform Block) 등을 의미할 수 있다.

[0076] - 유닛은, 블록과 구분하여 지칭하기 위해, 루마(luma) 성분 블록 및 이에 대응하는 크로마(chroma) 성분 블록, 그리고 각 블록에 대한 구문 요소(syntax element)를 포함하는 정보를 의미할 수 있다.

[0077] - 유닛의 크기 및 형태는 다양할 수 있다. 또한, 유닛은 다양한 크기 및 다양한 형태를 가질 수 있다. 특히 유닛의 형태는 정사각형뿐만 아니라 직사각형, 사다리꼴, 삼각형 및 오각형 등 2차원으로 표현될 수 있는 기하학적 도형을 포함할 수 있다.

[0078] - 또한, 유닛 정보는 유닛의 타입, 유닛의 크기, 유닛의 깊이, 유닛의 부호화 순서 및 유닛의 복호화 순서 등 중 적어도 하나 이상을 포함할 수 있다. 예를 들면, 유닛의 타입은 CU, PU, 잔차 유닛 및 TU 등 중 하나를 가리킬 수 있다.

[0079] - 하나의 유닛은 유닛에 비해 더 작은 크기를 갖는 하위 유닛으로 더 분할될 수 있다.

[0080] 깊이(depth): 깊이는 유닛의 분할된 정도를 의미할 수 있다. 또한, 유닛 깊이는 유닛을 트리 구조로 표현했을 때 유닛이 존재하는 레벨을 나타낼 수 있다.

[0081] - 유닛 분할 정보는 유닛의 깊이에 관한 깊이를 포함할 수 있다. 깊이는 유닛이 분할되는 회수 및/또는 정도를 나타낼 수 있다.

[0082] - 트리 구조에서, 루트 노드(root node)의 깊이가 가장 얕고, 리프 노드(leaf node)의 깊이가 가장 깊다고 볼 수 있다.

[0083] - 하나의 유닛은 트리 구조(tree structure)에 기반하여 깊이 정보(depth)를 가지면서 계층적으로(hierarchically) 복수의 하위 유닛들로 분할될 수 있다. 말하자면, 유닛 및 상기의 유닛의 분할에 의해 생성된 하위 유닛은 노드 및 상기의 노드의 자식 노드에 각각 대응할 수 있다. 각각의 분할된 하위 유닛은 깊이를 가질 수 있다. 깊이는 유닛이 분할된 회수 및/또는 정도를 나타내므로, 하위 유닛의 분할 정보는 하위 유닛의 크기에 관한 정보를 포함할 수도 있다.

[0084] - 트리 구조에서, 가장 상위 노드는 분할되지 않은 최초의 유닛에 대응할 수 있다. 가장 상위 노드는 루트 노드

로 칭해질 수 있다. 또한, 가장 상위 노드는 최소의 깊이 값을 가질 수 있다. 이 때, 가장 상위 노드는 레벨 0의 깊이를 가질 수 있다.

[0085] - 레벨 1의 깊이를 갖는 노드는 최초의 유닛이 한 번 분할됨에 따라 생성된 유닛을 나타낼 수 있다. 레벨 2의 깊이를 갖는 노드는 최초의 유닛이 두 번 분할됨에 따라 생성된 유닛을 나타낼 수 있다.

[0086] - 레벨 n의 깊이를 갖는 노드는 최초의 유닛이 n번 분할됨에 따라 생성된 유닛을 나타낼 수 있다.

[0087] - 리프 노드는 가장 하위의 노드일 수 있으며, 더 분할될 수 없는 노드일 수 있다. 리프 노드의 깊이는 최대 레벨일 수 있다. 예를 들면, 최대 레벨의 기정의된 값은 3일 수 있다.

[0088] - QT 깊이는 쿼드 분할에 대한 깊이를 나타낼 수 있다. BT 깊이는 이진 분할에 대한 깊이를 나타낼 수 있다. TT 깊이는 삼진 분할에 대한 깊이를 나타낼 수 있다.

[0089] 샘플(sample): 샘플은 블록을 구성하는 기반(base) 단위일 수 있다. 샘플은 비트 깊이(bit depth; Bd)에 따라서 0부터 2^{Bd-}1까지의 값들로서 표현될 수 있다.

[0090] - 샘플은 픽셀 또는 픽셀 값일 수 있다.

[0091] - 이하에서, 용어들 "픽셀", "화소" 및 "샘플"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0092] 코딩 트리 유닛(Coding Tree Unit; CTU): CTU는 하나의 루마 성분(Y) 코딩 트리 블록과, 상기의 루마 성분 코딩 트리 블록에 관련된 두 크로마 성분(Cb, Cr) 코딩 트리 블록들로 구성될 수 있다. 또한, CTU는 상기의 블록들과 상기의 블록들의 각 블록에 대한 구문 요소를 포함한 것을 의미할 수도 있다.

[0093] - 각 코딩 트리 유닛은 코딩 유닛, 예측 유닛 및 변환 유닛 등의 하위 유닛을 구성하기 위하여 쿼드 트리(Quad Tree: QT), 이진 트리(Binary Tree; BT) 및 삼진 트리(Ternary Tree; TT) 등과 같은 하나 이상의 분할 방식을 이용하여 분할될 수 있다. 또한, 각 코딩 트리 유닛은 하나 이상의 분할 방식들을 사용하는 복수 트리(MultiType Tree; MTT)을 이용하여 분할될 수 있다.

[0094] - CTU는 입력 영상의 분할에서와 같이, 영상의 복호화 및 부호화 과정에서의 처리 단위인 픽셀 블록을 지칭하기 위한 용어로서 사용될 수 있다.

[0095] 코딩 트리 블록(Coding Tree Block; CTB): 코딩 트리 블록은 Y 코딩 트리 블록, Cb 코딩 트리 블록, Cr 코딩 트리 블록 중 어느 하나를 지칭하기 위한 용어로 사용될 수 있다.

[0096] 주변 블록(neighbor block): 주변 블록은 대상 블록에 인접한 블록을 의미할 수 있다. 주변 블록은 복원된 주변 블록을 의미할 수도 있다.

[0097] - 이하에서, 용어들 "주변 블록" 및 "인접 블록(adjacent block)"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0098] 공간적 주변 블록(spatial neighbor block): 공간적 주변 블록은 대상 블록에 공간적으로 인접한 블록일 수 있다. 주변 블록은 공간적 주변 블록을 포함할 수 있다.

[0099] - 대상 블록 및 공간적 주변 블록은 대상 픽처 내에 포함될 수 있다.

[0100] - 공간적 주변 블록은 대상 블록에 경계가 맞닿은 블록 또는 대상 블록으로부터 소정의 거리 내에 위치한 블록을 의미할 수 있다.

[0101] - 공간적 주변 블록은 대상 블록의 꼭지점에 인접한 블록을 의미할 수 있다. 여기에서, 대상 블록의 꼭지점에 인접한 블록이란, 대상 블록에 가로로 인접한 이웃 블록에 세로로 인접한 블록 또는 대상 블록에 세로로 인접한 이웃 블록에 가로로 인접한 블록일 수 있다.

[0102] 시간적 주변 블록(temporal neighbor block): 시간적 주변 블록은 대상 블록에 시간적으로 인접한 블록일 수 있다. 주변 블록은 시간적 주변 블록을 포함할 수 있다.

[0103] - 시간적 주변 블록은 콜 블록(co-located block; col block)을 포함할 수 있다.

[0104] - 콜 블록은 이미 복원된 콜 픽처(co-located picture; col picture) 내의 블록일 수 있다. 콜 블록의 콜 픽처 내에서의 위치는 대상 블록의 대상 픽처 내의 위치에 대응할 수 있다. 또는, 콜 블록의 콜 픽처 내에서의 위치

는 대상 블록의 대상 픽처 내의 위치와 동일할 수 있다. 콜 픽처는 참조 픽처 리스트에 포함된 픽처일 수 있다.

[0105] - 시간적 주변 블록은 대상 블록의 공간적 주변 블록에 시간적으로 인접한 블록일 수 있다.

[0106] 예측 유닛(prediction unit): 인터 예측, 인트라 예측, 인터 보상(compensation), 인트라 보상 및 움직임 보상 등의 예측에 대한 기반 단위를 의미할 수 있다.

[0107] - 하나의 예측 유닛은 더 작은 크기를 갖는 복수의 파티션(partition)들 또는 하위 예측 유닛들로 분할될 수도 있다. 복수의 파티션들 또한 예측 또는 보상의 수행에 있어서의 기반 단위일 수 있다. 예측 유닛의 분할에 의해 생성된 파티션 또한 예측 유닛일 수 있다.

[0108] 예측 유닛 파티션(prediction unit partition): 예측 유닛 파티션은 예측 유닛이 분할된 형태를 의미할 수 있다.

[0109] 재구축된 이웃 유닛(reconstructed neighboring unit): 재구축된 이웃 유닛은 대상 유닛의 주변에 이미 복호화되어 재구축된 유닛일 수 있다.

[0110] - 재구축된 이웃 유닛은 대상 유닛에 대한 공간적(spatial) 인접 유닛 또는 시간적(temporal) 인접 유닛일 수 있다.

[0111] - 재구축된 공간적 주변 유닛은 대상 픽처 내의 유닛이면서 부호화 및/또는 복호화를 통해 이미 재구축된 유닛일 수 있다.

[0112] - 재구축된 시간적 주변 유닛은 참조 영상 내의 유닛이면서 부호화 및/또는 복호화를 통해 이미 재구축된 유닛일 수 있다. 재구축된 시간적 주변 유닛의 참조 영상 내에서의 위치는 대상 유닛의 대상 픽처 내에서의 위치와 같거나, 대상 유닛의 대상 픽처 내에서의 위치에 대응할 수 있다.

[0113] 파라미터 세트(parameter set): 파라미터 세트는 비트스트림 내의 구조(structure) 중 헤더(header) 정보에 해당할 수 있다. 예를 들면, 파라미터 세트는 비디오 파라미터 세트(Video Parameter Set; VPS), 시퀀스 파라미터 세트(Sequence Parameter Set: SPS), 픽처 파라미터 세트(Picture Parameter Set; PPS) 및 적응 파라미터 세트(Adaptation Parameter Set; APS) 등을 포함할 수 있다.

[0114] 또한, 파라미터 세트는 슬라이스(slice) 헤더 정보 및 타일 헤더 정보를 포함할 수 있다.

[0115] 율-왜곡 최적화(rate-distortion optimization): 부호화 장치는 코딩 유닛의 크기, 예측 모드, 예측 유닛의 크기, 움직임 정보 및, 변환 유닛의 크기 등의 조합을 이용해서 높은 부호화 효율을 제공하기 위해 율-왜곡 최적화를 사용할 수 있다.

[0116] - 율-왜곡 최적화 방식은 상기의 조합들 중에서 최적의 조합을 선택하기 위해 각 조합의 율-왜곡 비용(rate-distortion cost)을 계산할 수 있다. 율-왜곡 비용은 아래의 수식 1을 이용하여 계산될 수 있다. 일반적으로 상기 율-왜곡 비용이 최소가 되는 조합이 율-왜곡 최적화 방식에 있어서의 최적의 조합으로 선택될 수 있다.

[0117] [수식 1]

[0118] $D+\lambda*R$

[0119] - D는 왜곡을 나타낼 수 있다. D는 변환 유닛 내에서 원래의 변환 계수들 및 재구축된 변환 계수들 간의 차이 값들의 제곱들의 평균(mean square error)일 수 있다.

[0120] - R은 율을 나타낼 수 있다. R은 관련된 문맥 정보를 이용한 비트 율을 나타낼 수 있다.

[0121] - λ는 라그랑지안 승수(Lagrangian multiplier)를 나타낼 수 있다. R은 예측 모드, 움직임 정보 및 코드된 블록 플래그(coded block flag) 등과 같은 코딩 파라미터 정보뿐만 아니라, 변환 계수의 부호화에 의해 발생하는 비트도 포함할 수 있다.

[0122] - 부호화 장치는 정확한 D 및 R을 계산하기 위해 인터 예측 및/또는 인트라 예측, 변환, 양자화, 엔트로피 부호화, 역양자화, 역변환 등의 과정들을 수행할 수 있다. 이러한 과정들은 부호화 장치에서의 복잡도를 크게 증가시킬 수 있다.

[0123] 비트스트림(bitstream): 비트스트림은 부호화된 영상 정보를 포함하는 비트의 열을 의미할 수 있다.

[0124] 파라미터 세트(parameter set): 파라미터 세트는 비트스트림 내의 구조(structure) 중 헤더(header) 정보에 해

당할 수 있다.

[0125] - 파라미터 세트는 비디오 파라미터 세트(video parameter set), 시퀀스 파라미터 세트(sequence parameter set), 픽처 파라미터 세트(picture parameter set) 및 적응 파라미터 세트(adaptation parameter set) 중 적어도 하나를 포함할 수 있다. 또한, 파라미터 세트는 슬라이스(slice) 헤더의 정보 및 타일(tile) 헤더의 정보를 포함할 수도 있다.

[0126] 파싱(parsing): 파싱은 비트스트림을 엔트로피 복호화하여 구문 요소(syntax element)의 값을 결정하는 것을 의미할 수 있다. 또는, 파싱은 엔트로피 복호화 자체를 의미할 수 있다.

[0127] 심볼(symbol): 부호화 대상 유닛 및/또는 복호화 대상 유닛의 구문 요소, 코딩 파라미터(coding parameter) 및 변환 계수(transform coefficient) 등 중 적어도 하나를 의미할 수 있다. 또한, 심볼은 엔트로피 부호화의 대상 또는 엔트로피 복호화의 결과를 의미할 수 있다.

[0128] 참조 픽처(reference picture): 참조 픽처는 인터 예측 또는 움직임 보상을 위하여 유닛이 참조하는 영상을 의미할 수 있다. 또는, 참조 픽처는 인터 예측 또는 움직임 보상을 위해 대상 유닛이 참조하는 참조 유닛을 포함하는 영상일 수 있다.

[0129] 이하, 용어 "참조 픽처" 및 "참조 영상"은 동일한 의미로 사용될 수 있으며, 서로 교체되어 사용될 수 있다.

[0130] 참조 픽처 리스트(reference picture list): 참조 픽처 리스트는 인터 예측 또는 움직임 보상에 사용되는 하나 이상의 참조 영상들을 포함하는 리스트일 수 있다.

[0131] - 참조 픽처 리스트의 종류는 리스트 조합(List Combined; LC), 리스트 0(List 0; L0), 리스트 1(List 1; L1), 리스트 2(List 2; L2) 및 리스트 3(List 3; L3) 등이 있을 수 있다.

[0132] - 인터 예측에는 하나 이상의 참조 픽처 리스트들이 사용될 수 있다.

[0133] 인터 예측 지시자(inter prediction indicator): 인터 예측 지시자는 대상 유닛에 대한 인터 예측의 방향을 가리킬 수 있다. 인터 예측은 단방향 예측 및 양방향 예측 등 중 하나일 수 있다. 또는, 인터 예측 지시자는 대상 유닛의 예측 유닛을 생성할 때 사용되는 참조 영상의 개수를 나타낼 수 있다. 또는, 인터 예측 지시자는 대상 유닛에 대한 인터 예측 혹은 움직임 보상을 위해 사용되는 예측 블록의 개수를 의미할 수 있다.

[0134] 참조 픽처 색인(reference picture index): 참조 픽처 색인은 참조 픽처 리스트에서 특정 참조 영상을 지시하는 색인일 수 있다.

[0135] 움직임 벡터(Motion Vector; MV): 움직임 벡터는 인터 예측 또는 움직임 보상에서 사용되는 2차원의 벡터일 수 있다. 움직임 벡터는 대상 영상 및 참조 영상 간의 오프셋을 의미할 수 있다.

[0136] - 예를 들면, MV는 (mv_x, mv_y)와 같은 형태로 표현될 수 있다. mv_x는 수평(horizontal) 성분을 나타낼 수 있고, mv_y 는 수직(vertical) 성분을 나타낼 수 있다.

[0137] 탐색 영역(search range): 탐색 영역은 인터 예측 중 MV에 대한 탐색이 이루어지는 2차원의 영역일 수 있다. 예를 들면, 탐색 영역의 크기는 MxN일 수 있다. M 및 N은 각각 양의 정수일 수 있다.

[0138] 움직임 벡터 후보(motion vector candidate): 움직임 벡터 후보는 움직임 벡터를 예측할 때 예측 후보인 블록 혹은 예측 후보인 블록의 움직임 벡터를 의미할 수 있다.

[0139] - 움직임 벡터 후보는 움직임 벡터 후보 리스트에 포함될 수 있다.

[0140] 움직임 벡터 후보 리스트(motion vector candidate list): 움직임 벡터 후보 리스트는 하나 이상의 움직임 벡터 후보들을 이용하여 구성된 리스트를 의미할 수 있다.

[0141] 움직임 벡터 후보 색인(motion vector candidate index): 움직임 벡터 후보 색인은 움직임 벡터 후보 리스트 내의 움직임 벡터 후보를 가리키는 지시자를 의미할 수 있다. 또는, 움직임 벡터 후보 색인은 움직임 벡터 예측기(motion vector predictor)의 색인(index)일 수 있다.

[0142] 움직임 정보(motion information): 움직임 정보는 움직임 벡터, 참조 픽처 색인 및 인터 예측 지시자(inter prediction indicator) 뿐만 아니라 참조 픽처 리스트 정보, 참조 영상, 움직임 벡터 후보, 움직임 벡터 후보 색인, 머지 후보 및 머지 색인 등 중 적어도 하나를 포함하는 정보를 의미할 수 있다.

[0143] 머지 후보 리스트(merge candidate list): 머지 후보 리스트는 머지 후보를 이용하여 구성된 리스트를 의미할 수 있다.

[0144] 머지 후보(merge candidate): 머지 후보는 공간적 머지 후보, 시간적 머지 후보, 조합된 머지 후보, 조합 양예측(combined bi-prediction) 머지 후보 및 제로 머지 후보 등을 의미할 수 있다. 머지 후보는 예측 타입 정보, 각 리스트에 대한 참조 픽처 색인 및 움직임 벡터 등의 움직임 정보를 포함할 수 있다.

[0145] 머지 색인(merge index): 머지 색인은 머지 후보 리스트 내의 머지 후보를 가리키는 지시자일 수 있다.

[0146] - 머지 색인은 대상 유닛에 공간적으로 인접한 재구축된 유닛 및 대상 유닛에 시간적으로 인접한 재구축된 유닛 중 머지 후보를 유도한 재구축된 유닛을 지시할 수 있다.

[0147] - 머지 색인은 머지 후보의 움직임 정보들 중 적어도 하나를 지시할 수 있다.

[0148] 변환 유닛(transform unit): 변환 유닛은 변환, 역변환, 양자화, 역양자화, 변환 계수 부호화 및 변환 계수 복호화 등과 같은 잔차 신호(residual signal) 부호화 및/또는 잔차 신호 복호화에 있어서의 기본 유닛일 수 있다. 하나의 변환 유닛은 더 작은 크기의 복수의 변환 유닛들로 분할될 수 있다.

[0149] 스케일링(scaling): 스케일링은 변환 계수 레벨에 인수를 곱하는 과정을 의미할 수 있다.

[0150] - 변환 계수 레벨에 대한 스케일링의 결과로서, 변환 계수가 생성될 수 있다. 스케일링은 역양자화(dequantization)로 칭해질 수도 있다.

[0151] 양자화 파라미터(Quantization Parameter; QP): 양자화 파라미터는 양자화에서 변환 계수에 대해 변환 계수 레벨(transform coefficient level)을 생성할 때 사용되는 값을 의미할 수 있다. 또는, 양자화 파라미터는 역양자화에서 변환 계수 레벨을 스케일링(scaling)함으로써 변환 계수를 생성할 때 사용되는 값을 의미할 수도 있다. 또는, 양자화 파라미터는 양자화 스탭 크기(step size)에 매핑된 값일 수 있다.

[0152] 델타 양자화 파라미터(delta quantization parameter): 델타 양자화 파라미터는 예측된 양자화 파라미터 및 대상 유닛의 양자화 파라미터의 차분(differential) 값을 의미한다.

[0153] 스캔(scan): 스캔은 유닛, 블록 또는 행렬 내의 계수들의 순서를 정렬하는 방법을 의미할 수 있다. 예를 들면, 2차원 배열을 1차원 배열 형태로 정렬하는 것을 스캔이라고 칭할 수 있다. 또는, 1차원 배열을 2차원 배열 형태로 정렬하는 것도 스캔 또는 역 스캔(inverse scan)이라고 칭할 수 있다.

[0154] 변환 계수(transform coefficient): 변환 계수는 부호화 장치에서 변환을 수행함에 따라 생성된 계수 값일 수 있다. 또는, 변환 계수는 복호화 장치에서 엔트로피 복호화 및 역양자화 중 적어도 하나를 수행함에 따라 생성된 계수 값일 수 있다.

[0155] - 변환 계수 또는 잔차 신호에 양자화를 적용함으로써 생성된 양자화된 레벨 또는 양자화된 변환 계수 레벨 또한 변환 계수의 의미에 포함될 수 있다.

[0156] 양자화된 레벨(quantized level): 양자화된 레벨은 부호화 장치에서 변환 계수 또는 잔차 신호에 양자화를 수행함으로써 생성된 값을 의미할 수 있다. 또는, 양자화된 레벨은 복호화 장치에서 역양자화를 수행함에 있어서 역양자화의 대상이 되는 값을 의미할 수도 있다.

[0157] - 변환 및 양자화의 결과인 양자화된 변환 계수 레벨도 양자화된 레벨의 의미에 포함될 수 있다.

[0158] 넌제로 변환 계수(non-zero transform coefficient): 넌제로 변환 계수는 0이 아닌 값을 갖는 변환 계수 또는 0이 아닌 값을 갖는 변환 계수 레벨을 의미할 수 있다. 또는, 넌제로 변환 계수는 값의 크기가 0이 아닌 변환 계수 또는 값의 크기가 0이 아닌 변환 계수 레벨을 의미할 수 있다.

[0159] 양자화 행렬(quantization matrix): 양자화 행렬은 영상의 주관적 화질 혹은 객관적 화질을 향상시키기 위해서 양자화 과정 또는 역양자화 과정에서 이용되는 행렬을 의미할 수 있다. 양자화 행렬은 스케일링 리스트(scaling list)라고도 칭해질 수 있다.

[0160] 양자화 행렬 계수(quantization matrix coefficient): 양자화 행렬 계수는 양자화 행렬 내의 각 원소(element)를 의미할 수 있다. 양자화 행렬 계수는 행렬 계수(matrix coefficient)라고도 칭해질 수 있다.

[0161] 디폴트 행렬(default matrix): 기본 행렬은 부호화 장치 및 복호화 장치에서 기정의된 양자화 행렬일 수 있다.

[0162] 비 디폴트 행렬(non-default matrix): 비 디폴트 행렬은 부호화 장치 및 복호화 장치에서 기정의되어 있지 않은

양자화 행렬일 수 있다. 비 디폴트 행렬은 부호화 장치로부터 복호화 장치로 시그널링될 수 있다.

[0163] 가장 가능성있는 모드(Most Probable Mode; MPM): MPM은 대상 블록의 인트라 예측을 위해 사용될 가능성이 높은 인트라 예측 모드를 나타낼 수 있다.

[0164] 부호화 장치 및 복호화 장치는 대상 블록에 관련된 코딩 파라미터 및 대상 블록에 관련된 개체의 속성에 기반하여 하나 이상의 MPM들을 결정할 수 있다.

[0165] 부호화 장치 및 복호화 장치는 참조 블록의 인트라 예측 모드에 기반하여 하나 이상의 MPM들을 결정할 수 있다. 참조 블록은 복수일 수 있다. 복수의 참조 블록들은 대상 블록의 좌측에 인접한 공간적 주변 블록 및 대상 블록의 상단에 인접한 공간적 주변 블록을 포함할 수 있다. 말하자면, 참조 블록들에 대하여 어떠한 인트라 예측 모드들이 사용되었는가에 따라서 서로 다른 하나 이상의 MPM들이 결정될 수 있다.

[0166] 하나 이상의 MPM들은 부호화 장치 및 복호화 장치에서 동일한 방식으로 결정될 수 있다. 말하자면, 부호화 장치 및 복호화 장치는 동일한 하나 이상의 MPM들을 포함하는 MPM 리스트를 공유할 수 있다.

[0167] MPM 리스트: MPM 리스트는 하나 이상의 MPM들을 포함하는 리스트일 수 있다. MPM 리스트 내의 하나 이상의 MPM들의 개수는 기정의될 수 있다.

[0168] MPM 지시자: MPM 지시자는 MPM 리스트의 하나 이상의 MPM들 중 대상 블록의 인트라 예측을 위해 사용되는 MPM을 지시할 수 있다. 예를 들면, MPM 지시자는 MPM 리스트에 대한 인덱스일 수 있다.

[0169] MPM 리스트는 부호화 장치 및 복호화 장치에서 동일한 방식으로 결정되기 때문에 MPM 리스트 자체는 부호화 장치로부터 복호화 장치로 전송될 필요가 없을 수 있다.

[0170] MPM 지시자는 부호화 장치로부터 복호화 장치로 시그널링될 수 있다. MPM 지시자가 시그널링됨에 따라 복호화 장치는 MPM 리스트의 MPM들 중 대상 블록에 대한 인트라 예측을 위해 사용될 MPM을 결정할 수 있다.

[0171] MPM 사용 지시자: MPM 사용 지시자는 대상 블록에 대한 예측을 위해 MPM 사용 모드가 사용될지 여부를 지시할 수 있다. MPM 사용 모드는 MPM 리스트를 사용하여 대상 블록에 대한 인트라 예측을 위해 사용될 MPM을 결정하는 모드일 수 있다.

[0172] MPM 사용 지시자는 부호화 장치로부터 복호화 장치로 시그널링될 수 있다.

[0173] 시그널링: 시그널링은 정보가 부호화 장치로부터 복호화 장치로 전송되는 것을 나타낼 수 있다. 또는, 시그널링은 정보를 비트스트림 또는 기록 매체 내에 포함시키는 것을 의미할 수 있다. 부호화 장치에 의해 시그널링된 정보는 복호화 장치에 의해 사용될 수 있다.

[0174] 도 1은 본 발명이 적용되는 부호화 장치의 일 실시예에 따른 구성을 나타내는 블록도이다.

[0175] 부호화 장치(100)는 인코더, 비디오 부호화 장치 또는 영상 부호화 장치일 수 있다. 비디오는 하나 이상의 영상들을 포함할 수 있다. 부호화 장치(100)는 비디오의 하나 이상의 영상들을 순차적으로 부호화할 수 있다.

[0176] 도 1을 참조하면, 부호화 장치(100)는 인터 예측부(110), 인트라 예측부(120), 스위치(115), 감산기(125), 변환부(130), 양자화부(140), 엔트로피 부호화부(150), 역양자화부(160), 역변환부(170), 가산기(175), 필터부(180) 및 참조 픽처 버퍼(190)를 포함할 수 있다.

[0177] 부호화 장치(100)는 인트라 모드 및/또는 인터 모드를 사용하여 대상 영상에 대한 부호화를 수행할 수 있다.

[0178] 또한, 부호화 장치(100)는 대상 영상에 대한 부호화를 통해 부호화의 정보를 포함하는 비트스트림을 생성할 수 있고, 생성된 비트스트림을 출력할 수 있다. 생성된 비트스트림은 컴퓨터 판독가능한 기록 매체에 저장될 수 있고, 유/무선 전송 매체를 통해 스트리밍될 수 있다.

[0179] 예측 모드로서, 인트라 모드가 사용되는 경우, 스위치(115)는 인트라로 전환될 수 있다. 예측 모드로서, 인터 모드가 사용되는 경우, 스위치(115)는 인터로 전환될 수 있다.

[0180] 부호화 장치(100)는 대상 블록에 대한 예측 블록을 생성할 수 있다. 또한, 부호화 장치(100)는 예측 블록이 생성된 후, 대상 블록 및 예측 블록의 차분(residual)을 부호화할 수 있다.

[0181] 예측 모드가 인트라 모드인 경우, 인트라 예측부(120)는 대상 블록의 주변에 있는, 이미 부호화/복호화된 블록의 픽셀을 참조 샘플로서 이용할 수 있다. 인트라 예측부(120)는 참조 샘플을 이용하여 대상 블록에 대한 공간적 예측을 수행할 수 있고, 공간적 예측을 통해 대상 블록에 대한 예측 샘플들을 생성할 수 있다.

[0182] 인터 예측부(110)는 움직임 예측부 및 움직임 보상부를 포함할 수 있다.

[0183] 예측 모드가 인터 모드인 경우, 움직임 예측부는, 움직임 예측 과정에서 참조 영상으로부터 대상 블록과 가장 매치가 잘 되는 영역을 검색할 수 있고, 검색된 영역을 이용하여 대상 블록 및 검색된 영역에 대한 움직임 벡터를 도출할 수 있다.

[0184] 참조 영상은 참조 픽처 버퍼(190)에 저장될 수 있으며, 참조 영상에 대한 부호화 및/또는 복호화가 처리되었을 때 참조 픽처 버퍼(190)에 저장될 수 있다.

[0185] 움직임 보상부는 움직임 벡터를 이용하는 움직임 보상을 수행함으로써 대상 블록에 대한 예측 블록을 생성할 수 있다. 여기에서, 움직임 벡터는 인터 예측에 사용되는 2차원 벡터일 수 있다. 또한 움직임 벡터는 대상 영상 및 참조 영상 간의 오프셋(offset)을 나타낼 수 있다.

[0186] 움직임 예측부 및 움직임 보상부는 움직임 벡터가 정수가 아닌 값을 가진 경우 참조 영상 내의 일부 영역에 대해 보간 필터(interpolation filter)를 적용함으로써 예측 블록을 생성할 수 있다. 인터 예측 또는 움직임 보상을 수행하기 위해, CU를 기준으로 CU에 포함된 PU의 움직임 예측 및 움직임 보상의 방법이 스킵 모드(skip mode), 머지 모드(merge mode), 향상된 움직임 벡터 예측(advanced motion vector prediction; AMVP) 모드 및 현재 픽처 참조 모드 중 어떠한 방법인지 여부가 판단될 수 있고, 각 모드에 따라 인터 예측 또는 움직임 보상이 수행될 수 있다.

[0187] 감산기(125)는 대상 블록 및 예측 블록의 차분인 잔차 블록(residual block)을 생성할 수 있다. 잔차 블록은 잔차 신호로 칭해질 수도 있다.

[0188] 잔차 신호는 원 신호 및 예측 신호 간의 차이(difference)를 의미할 수 있다. 또는, 잔차 신호는 원신호 및 예측 신호 간의 차이를 변환(transform)하거나 양자화하거나 또는 변환 및 양자화함으로써 생성된 신호일 수 있다. 잔차 블록은 블록 단위에 대한 잔차 신호일 수 있다.

[0189] 변환부(130)는 잔차 블록에 대해 변환(transform)을 수행하여 변환 계수를 생성할 수 있고, 생성된 변환 계수(transform coefficient)를 출력할 수 있다. 여기서, 변환 계수는 잔차 블록에 대한 변환을 수행함으로써 생성된 계수 값일 수 있다.

[0190] 변환부(130)는 변환을 수행함에 있어서 기정의된 복수의 변환 방법들 중 하나를 사용할 수 있다.

[0191] 기정의된 복수의 변환 방법들은 이산 코사인 변환(Discrete Cosine Transform; DCT), 이산 사인 변환(Discrete Sine Transform; DST) 및 카루넨-루베 변환(Karhunen-Loeve Transform; KLT) 기반 변환 등을 포함할 수 있다.

[0192] 잔차 블록에 대한 변환을 위해 사용되는 변환 방법은 대상 블록 및/또는 주변 블록에 대한 코딩 파라미터들 중 적어도 하나에 따라 결정될 수 있다. 예를 들면, 변환 방법은 PU에 대한 인터 예측 모드, PU에 대한 인트라 예측 모드, TU의 크기 및 TU의 형태 중 적어도 하나에 기반하여 결정될 수 있다. 또는, 변환 방법을 지시하는 변환 정보가 부호화 장치(100)로부터 복호화 장치(200)로 시그널링될 수도 있다.

[0193] 변환 스킵(transform skip) 모드가 적용되는 경우, 변환부(130)는 잔차 블록에 대한 변환을 생략할 수도 있다.

[0194] 변환 계수에 양자화를 적용함으로써 양자화된 변환 계수 레벨(transform coefficient level) 또는 양자화된 레벨이 생성될 수 있다. 이하, 실시예들에서는 양자화된 변환 계수 레벨 및 양자화된 레벨도 변환 계수로 칭해질 수 있다.

[0195] 양자화부(140)는 변환 계수를 양자화 파라미터에 맞춰 양자화함으로써 양자화된 변환 계수 레벨(quantized transform coefficient level)(말하자면, 양자화된 레벨 또는 양자화된 계수)를 생성할 수 있다. 양자화부(140)는 생성된 양자화된 변환 계수 레벨을 출력할 수 있다. 이때, 양자화부(140)에서는 양자화 행렬을 사용하여 변환 계수를 양자화할 수 있다.

[0196] 엔트로피 부호화부(150)는, 양자화부(140)에서 산출된 값들 및/또는 부호화 과정에서 산출된 코딩 파라미터 값들 등에 기초하여 확률 분포에 따른 엔트로피 부호화를 수행함으로써 비트스트림(bitstream)을 생성할 수 있다. 엔트로피 부호화부(150)는 생성된 비트스트림을 출력할 수 있다.

[0197] 엔트로피 부호화부(150)는 영상의 픽셀에 관한 정보 및 영상의 복호화를 위한 정보에 대한 엔트로피 부호화를 수행할 수 있다. 예를 들면, 영상의 복호화를 위한 정보는 구문 요소(syntax element) 등을 포함할 수 있다.

[0198] 엔트로피 부호화가 적용되는 경우, 높은 발생 확률을 갖는 심볼에 적은 수의 비트가 할당될 수 있고, 낮은 발생

확률을 갖는 심볼에 많은 수의 비트가 할당될 수 있다. 이러한 할당을 통해 심볼이 표현됨에 따라, 부호화의 대상인 심볼들에 대한 비트열(bitstring)의 크기가 감소될 수 있다. 따라서, 엔트로피 부호화를 통해서 영상 부호화의 압축 성능이 향상될 수 있다.

[0199] 또한, 엔트로피 부호화부(150)는 엔트로피 부호화를 위해 지수 골롬(exponential golomb), 문맥-적응형 가변 길이 코딩(Context-Adaptive Variable Length Coding; CAVLC) 및 문맥-적응형 이진 산술 코딩(Context-Adaptive Binary Arithmetic Coding; CABAC) 등과 같은 부호화 방법을 사용할 수 있다. 예를 들면, 엔트로피 부호화부(150)는 가변 길이 부호화(Variable Length Coding/Code; VLC) 테이블을 이용하여 엔트로피 부호화를 수행할 수 있다. 예를 들면, 엔트로피 부호화부(150)는 대상 심볼에 대한 이진화(binarization) 방법을 도출할 수 있다. 또한, 엔트로피 부호화부(150)는 대상 심볼/빈(bin)의 확률 모델(probability model)을 도출할 수 있다. 엔트로피 부호화부(150)는 도출된 이진화 방법, 확률 모델 및 문맥 모델(context model)을 사용하여 산술 부호화를 수행할 수도 있다.

[0200] 엔트로피 부호화부(150)는 양자화된 변환 계수 레벨을 부호화하기 위해 변환 계수 스캐닝(transform coefficient scanning) 방법을 통해 2차원의 블록의 형태(form)의 계수를 1차원의 벡터의 형태로 변경할 수 있다.

[0201] 코딩 파라미터는 부호화 및/또는 복호화를 위해 요구되는 정보일 수 있다. 코딩 파라미터는 부호화 장치(100)에서 부호화되어 부호화 장치(100)로부터 복호화 장치로 전달되는 정보를 포함할 수 있고, 부호화 혹은 복호화 과정에서 유추될 수 있는 정보를 포함할 수 있다. 예를 들면, 복호화 장치로 전달되는 정보로서, 구문 요소가 있다.

[0202] 코딩 파라미터(coding parameter)는 구문 요소와 같이 부호화 장치에서 부호화되고, 부호화 장치로부터 복호화 장치로 시그널링되는 정보(또는, 플래그, 인덱스 등)뿐만 아니라, 부호화 과정 또는 복호화 과정에서 유도되는 정보를 포함할 수 있다. 또한, 코딩 파라미터는 영상을 부호화하거나 복호화함에 있어서 요구되는 정보를 포함할 수 있다. 예를 들면, 유닛/블록의 크기, 유닛/블록의 깊이, 유닛/블록의 분할 정보, 유닛/블록의 분할 구조, 유닛/블록이 쿼드 트리 형태로 분할되는지 여부를 나타내는 정보, 유닛/블록이 이진 트리 형태로 분할되는지 여부를 나타내는 정보, 이진 트리 형태의 분할 방향(가로 방향 또는 세로 방향), 이진 트리 형태의 분할 형태(대칭 분할 또는 비대칭 분할), 유닛/블록이 삼진 트리 형태로 분할되는지 여부를 나타내는 정보, 삼진 트리 형태의 분할 방향(가로 방향 또는 세로 방향), 삼진 트리 형태의 분할 형태(대칭 분할 또는 비대칭 분할 등), 유닛/블록이 복합 트리 형태로 분할되는지 여부를 나타내는 정보, 복합 트리 형태의 분할의 조합 및 방향(가로 방향 또는 세로 방향 등), 예측 방식(인트라 예측 또는 인터 예측), 인트라 예측 모드/방향, 참조 샘플 필터링 방법, 예측 블록 필터링 방법, 예측 블록 경계 필터링 방법, 필터링의 필터 탭, 필터링의 필터 계수, 인터 예측 모드, 움직임 정보, 움직임 벡터, 참조 픽처 색인, 인터 예측 방향, 인터 예측 지시자, 참조 픽처 리스트, 참조 영상, 움직임 벡터 예측기, 움직임 벡터 예측 후보, 움직임 벡터 후보 리스트, 머지 모드를 사용하는지 여부를 나타내는 정보, 머지 후보, 머지 후보 리스트, 스킵(skip) 모드를 사용하는지 여부를 나타내는 정보, 보간 필터의 종류, 보간 필터의 필터 탭, 보간 필터의 필터 계수, 움직임 벡터 크기, 움직임 벡터 표현 정확도, 변환 종류, 변환 크기, 1차 변환을 사용하는지 여부를 나타내는 정보, 추가(2차) 변환을 사용하는지 여부를 나타내는 정보, 1차 변환 선택 정보(또는, 1차 변환 인덱스), 2차 변환 선택 정보(또는, 2차 변환 인덱스), 잔차 신호의 유무를 나타내는 정보, 코드된 블록 패턴(coded block pattern), 코드된 블록 플래그(coded block flag), 양자화 파라미터, 양자화 행렬, 인트라-루프 필터에 대한 정보, 인트라-루프 필터를 적용하는지 여부에 대한 정보, 인트라-루프 필터의 계수, 인트라-루프의 필터 탭, 인트라 루프 필터의 모양(shape)/형태(form), 디블록킹 필터를 적용하는지 여부를 나타내는 정보, 디블록킹 필터 계수, 디블록킹 필터 탭, 디블록킹 필터 강도, 디블록킹 필터 모양/형태, 적응적 샘플 오프셋을 적용하는지 여부를 나타내는 정보, 적응적 샘플 오프셋 값, 적응적 샘플 오프셋 카테고리, 적응적 샘플 오프셋 종류, 적응적 루프-내(in-loop) 필터를 적용하는지 여부, 적응적 루프-내 필터 계수, 적응적 루프-내 필터 탭, 적응적 루프-내 필터 모양/형태, 이진화/역이진화 방법, 문맥 모델, 문맥 모델 결정 방법, 문맥 모델 업데이트 방법, 레귤러 모드를 수행하는지 여부, 바이패스 모드를 수행하는지 여부, 문맥 빈, 바이패스 빈, 변환 계수, 변환 계수 레벨, 변환 계수 레벨 스캐닝 방법, 영상의 디스플레이/출력 순서, 슬라이스 식별 정보, 슬라이스 타입, 슬라이스 분할 정보, 타일 식별 정보, 타일 타입, 타일 분할 정보, 픽처 타입, 비트 심도, 루마 신호에 대한 정보 및 크로마 신호에 대한 정보 중 적어도 하나의 값, 조합된 형태 또는 통계가 코딩 파라미터에 포함될 수 있다. 예측 방식은 인트라 예측 모드 및 인터 예측 모드 중 하나의 예측 모드를 나타낼 수 있다.

[0203] 1차 변환 선택 정보는 대상 블록에 적용되는 1차 변환을 나타낼 수 있다.

[0204] 2차 변환 선택 정보는 대상 블록에 적용되는 2차 변환을 나타낼 수 있다.

[0205] 잔차 신호는 원 신호 및 예측 신호 간의 차분(difference)을 나타낼 수 있다. 또는, 잔차 신호는 원신호 및 예측 신호 간의 차분을 변환(transform)함으로써 생성된 신호일 수 있다. 또는, 잔차 신호는 원 신호 및 예측 신호 간의 차분을 변환 및 양자화함으로써 생성된 신호일 수 있다. 잔차 블록은 블록에 대한 잔차 신호일 수 있다.

[0206] 여기서, 플래그 또는 인덱스를 시그널링(signaling)한다는 것은 부호화 장치(100)에서는 플래그 또는 인덱스에 대한 엔트로피 부호화(entropy encoding)를 수행함으로써 생성된 엔트로피 부호화된 플래그 또는 엔트로피 부호화된 인덱스를 비트스트림(Bitstream)에 포함시키는 것을 의미할 수 있고, 복호화 장치(200)에서는 비트스트림으로부터 추출된 엔트로피 부호화된 플래그 또는 엔트로피 부호화된 인덱스에 대한 엔트로피 복호화(entropy decoding)를 수행함으로써 플래그 또는 인덱스를 획득하는 것을 의미할 수 있다.

[0207] 부호화 장치(100)에 의해 인터 예측을 통한 부호화가 수행되기 때문에, 부호화된 대상 영상은 이후에 처리되는 다른 영상(들)에 대하여 참조 영상으로서 사용될 수 있다. 따라서, 부호화 장치(100)는 부호화된 대상 영상을 다시 재구축 또는 복호화할 수 있고, 재구축 또는 복호화된 영상을 참조 영상으로서 참조 픽처 버퍼(190)에 저장할 수 있다. 복호화를 위해 부호화된 대상 영상에 대한 역양자화 및 역변환이 처리될 수 있다.

[0208] 양자화된 레벨은 역양자화부(160)에서 역양자화될(inversely quantized) 수 있고, 역변환부(170)에서 역변환될(inversely transformed) 수 있다. 역양자화부(160)는 양자화된 레벨에 대한 역양자화를 수행함으로써 역양자화된 계수를 생성할 수 있다. 역변환부(170)는 역양자화된 계수에 대한 역변환을 수행함으로써 재구축된 잔차 블록이 생성될 수 있다. 말하자면, 재구축된 잔차 블록은 역양자화 및 역변환된 계수일 수 있다.

[0209] 역양자화 및 역변환된 계수는 가산기(175)를 통해 예측 블록과 합해질 수 있다, 역양자화 및/또는 역변환된 계수와 예측 블록을 합함으로써 재구축된(reconstructed) 블록이 생성될 수 있다. 여기서, 역양자화 및/또는 역변환된 계수는 역양자화(dequantization) 및 역변환(inverse-transformation) 중 적어도 하나 이상이 수행된 계수를 의미할 수 있고, 재구축된 잔차 블록을 의미할 수 있다.

[0210] 재구축된 블록은 필터부(180)를 거칠 수 있다. 필터부(180)는 디블록킹 필터(deblocking filter), 샘플 적응적 오프셋(Sample Adaptive Offset; SAO), 적응적 루프 필터(Adaptive Loop Filter; ALF) 및 논 로컬 필터(Non Local Filter; NLF) 중 적어도 하나 이상을 재구축된 블록 또는 재구축된 픽처에 적용할 수 있다. 필터부(180)는 루프-내(in-loop) 필터로 칭해질 수도 있다.

[0211] 디블록킹 필터는 블록들 간의 경계에서 발생한 블록 왜곡을 제거할 수 있다. 디블록킹 필터를 적용할지 여부를 판단하기 위해, 블록에 포함된 몇 개의 열 또는 행에 포함된 픽셀(들)에 기반하여 대상 블록에 디블록킹 필터를 적용할지 여부가 판단될 수 있다.

[0212] 대상 블록에 디블록킹 필터를 적용하는 경우, 적용되는 필터는 요구되는 디블록킹 필터링의 강도에 따라 다를 수 있다. 말하자면, 서로 다른 필터들 중 디블록킹 필터링의 강도에 따라 결정된 필터가 대상 블록에 적용될 수 있다. 대상 블록에 디블록킹 필터가 적용되는 경우, 요구되는 디블록킹 필터링의 강도에 따라 강한 필터(strong filter) 및 약한 필터(weak filter) 중 하나의 필터가 대상 블록에 적용될 수 있다.

[0213] 또한, 대상 블록에 수직 방향 필터링 및 수평 방향 필터링이 수행되는 경우, 수평 방향 필터링 및 수직 방향 필터링이 병행으로 처리될 수 있다.

[0214] SAO는 코딩 에러에 대한 보상을 위해 픽셀의 픽셀 값에 적정한 오프셋(offset)을 더할 수 있다. SAO는 디블록킹이 적용된 영상에 대해, 픽셀의 단위로 원본 영상 및 디블록킹이 적용된 영상 간의 차이에 대하여 오프셋을 사용하는 보정을 수행할 수 있다. 영상에 대한 오프셋 보정을 수행하기 위해, 영상에 포함된 픽셀들을 일정한 수의 영역들로 구분한 후, 구분된 영역들 중 오프셋이 수행될 영역을 결정하고 결정된 영역에 오프셋을 적용하는 방법이 사용될 수 있고, 영상의 각 픽셀의 에지 정보를 고려하여 오프셋을 적용하는 방법이 사용될 수 있다.

[0215] ALF는 재구축된 영상 및 원래의 영상을 비교한 값에 기반하여 필터링을 수행할 수 있다. 영상에 포함된 픽셀들을 소정의 그룹들로 분할한 후, 각 분할된 그룹에 적용될 필터가 결정될 수 있고, 그룹 별로 차별적으로 필터링이 수행될 수 있다. 루마 신호에 대하여, 적응적 루프 필터를 적용할지 여부에 관련된 정보는 CU 별로 시그널링될 수 있다. 각 블록에 적용될 ALF 의 모양 및 필터 계수는 블록 별로 다를 수 있다. 또는, 블록의 특징과는 무

관하게, 고정된 형태의 ALF가 블록에 적용될 수 있다.

[0216] 논 로컬 필터는 대상 블록과 유사한 재구축된 블록들에 기반하여 필터링을 수행할 수 있다. 재구축된 영상에서 대상 블록과 유사한 영역이 선택될 수 있고, 선택된 유사한 영역의 통계적 성질을 사용하여 대상 블록의 필터링이 수행될 수 있다. 논 로컬 필터를 적용할지 여부에 관련된 정보는 CU에 대하여 시그널링될 수 있다. 또한, 블록들에 적용될 논 로컬 필터의 모양들 및 필터 계수들은 블록에 따라서 서로 다를 수 있다.

[0217] 필터부(180)를 거친 재구축된 블록 또는 재구축된 영상은 참조 픽처 버퍼(190)에 저장될 수 있다. 필터부(180)를 거친 재구축된 블록은 참조 픽처의 일부일 수 있다. 말하자면, 참조 픽처는 필터부(180)를 거친 재구축된 블록들로 구성된 재구축된 픽처일 수 있다. 저장된 참조 픽처는 이후 인터 예측에 사용될 수 있다.

[0218] 도 2는 본 발명이 적용되는 복호화 장치의 일 실시예에 따른 구성을 나타내는 블록도이다.

[0219] 복호화 장치(200)는 디코더, 비디오 복호화 장치 또는 영상 복호화 장치일 수 있다.

[0220] 도 2를 참조하면, 복호화 장치(200)는 엔트로피 복호화부(210), 역양자화부(220), 역변환부(230), 인트라 예측부(240), 인터 예측부(250), 스위치(245), 가산기(255), 필터부(260) 및 참조 픽처 버퍼(270)를 포함할 수 있다.

[0221] 복호화 장치(200)는 부호화 장치(100)에서 출력된 비트스트림을 수신할 수 있다. 복호화 장치(200)는 컴퓨터 판독가능한 기록 매체에 저장된 비트스트림을 수신할 수 있고, 유/무선 전송 매체를 통해 스트리밍되는 비트스트림을 수신할 수 있다.

[0222] 복호화 장치(200)는 비트스트림에 대하여 인트라 모드 및/또는 인터 모드의 복호화를 수행할 수 있다. 또한, 복호화 장치(200)는 복호화를 통해 재구축된 영상 또는 복호화된 영상을 생성할 수 있고, 생성된 재구축된 영상 또는 복호화된 영상을 출력할 수 있다.

[0223] 예를 들면, 복호화에 사용되는 예측 모드에 따른 인트라 모드 또는 인터 모드로의 전환은 스위치(245)에 의해 이루어질 수 있다. 복호화에 사용되는 예측 모드가 인트라 모드인 경우 스위치(245)가 인트라로 전환될 수 있다. 복호화에 사용되는 예측 모드가 인터 모드인 경우 스위치(245)가 인터로 전환될 수 있다.

[0224] 복호화 장치(200)는 입력된 비트스트림을 복호화함으로써 재구축된 잔차 블록(reconstructed residual block)을 획득할 수 있고, 예측 블록을 생성할 수 있다. 재구축된 잔차 블록 및 예측 블록이 획득되면, 복호화 장치(200)는 재구축된 잔차 블록 및 예측 블록을 더함으로써 복호화의 대상이 되는 재구축된 블록을 생성할 수 있다.

[0225] 엔트로피 복호화부(210)는 비트스트림에 대한 확률 분포에 기초하여 비트스트림에 대한 엔트로피 복호화를 수행함으로써 심볼들을 생성할 수 있다. 생성된 심볼들은 양자화된 변환 계수 레벨(quantized transform coefficient level)(말하자면, 양자화된 레벨 또는 양자화된 계수) 형태의 심볼을 포함할 수 있다. 여기에서, 엔트로피 복호화 방법은 상술된 엔트로피 부호화 방법과 유사할 수 있다. 예를 들면, 엔트로피 복호화 방법은 상술된 엔트로피 부호화 방법의 역과정일 수 있다.

[0226] 엔트로피 복호화부(210)는 양자화된 변환 계수 레벨을 복호화하기 위해 변환 계수 스캐닝 방법을 통해 1차원의 벡터의 형태의 계수를 2차원의 블록의 형태로 변경할 수 있다.

[0227] 예를 들면, 우상단 대각 스캔을 이용하여 블록의 계수들을 스캔함으로써 계수들이 2차원 블록 형태로 변경될 수 있다. 또는, 블록의 크기 및/또는 인트라 예측 모드에 따라 우상단 대각 스캔, 수직 스캔 및 수평 스캔 중 어떤 스캔이 사용될 것인지가 결정될 수 있다.

[0228] 양자화된 계수는 역양자화부(220)에서 역양자화될 수 있다. 역양자화부(220)는 양자화된 계수에 대한 역양자화를 수행함으로써 역양자화된 계수를 생성할 수 있다. 또한, 역양자화된 계수는 역변환부(230)에서 역변환될 수 있다. 역변환부(230)는 역양자화된 계수에 대한 역변환을 수행함으로써 재구축된 잔차 블록을 생성할 수 있다. 양자화된 계수에 대한 역양자화 및 역변환이 수행된 결과로서, 재구축된 잔차 블록이 생성될 수 있다. 이때, 역양자화부(220)는 재구축된 잔차 블록을 생성함에 있어서 양자화된 계수에 양자화 행렬을 적용할 수 있다.

[0229] 인트라 모드가 사용되는 경우, 인트라 예측부(240)는 대상 블록 주변의 이미 복호화된 블록의 픽셀 값을 이용하는 공간적 예측을 수행함으로써 예측 블록을 생성할 수 있다.

[0230] 인터 예측부(250)는 움직임 보상부를 포함할 수 있다. 또는, 인터 예측부(250)는 움직임 보상부로 명명될 수 있다.

[0231] 인터 모드가 사용되는 경우, 움직임 보상부는 움직임 벡터 및 참조 픽처 버퍼(270)에 저장된 참조 영상을 이용하는 움직임 보상을 수행함으로써 예측 블록을 생성할 수 있다.

[0232] 움직임 보상부는 움직임 벡터가 정수가 아닌 값을 가진 경우, 참조 영상 내의 일부 영역에 대해 보간 필터를 적용할 수 있고, 보간 필터가 적용된 참조 영상을 사용하여 예측 블록을 생성할 수 있다. 움직임 보상부는 움직임 보상을 수행하기 위해 CU를 기준으로 CU에 포함된 PU를 위해 사용되는 움직임 보상 방법이 스킵 모드, 머지 모드, AMVP 모드 및 현재 픽처 참조 모드 중 어떤 모드인가를 결정할 수 있고, 결정된 모드에 따라 움직임 보상을 수행할 수 있다.

[0233] 재구축된 잔차 블록 및 예측 블록은 가산기(255)를 통해 더해질 수 있다. 가산기(255)는 재구축된 잔차 블록 및 예측 블록을 더함으로써 재구축된 블록을 생성할 수 있다.

[0234] 재구축된 블록은 필터부(260)를 거칠 수 있다. 필터부(260)는 디블록킹 필터, SAO, ALF 및 논 로컬 필터 중 적어도 하나를 재구축된 블록 또는 재구축된 영상에 적용할 수 있다. 재구축된 영상은 재구축된 블록을 포함하는 픽처일 수 있다.

[0235] 필터부(260)를 거친 재구축된 영상은 부호화 장치(100)에 의해 출력될 수 있으며, 부호화 장치(100)에 의해 사용될 수 있다.

[0236] 필터부(260)를 거친 재구축된 영상은 참조 픽처 버퍼(270)에 참조 픽처로서 저장될 수 있다. 필터부(260)를 거친 재구축된 블록은 참조 픽처의 일부일 수 있다. 말하자면, 참조 픽처는 필터부(260)를 거친 재구축된 블록들로 구성된 영상일 수 있다. 저장된 참조 픽처는 이후 인터 예측을 위해 사용될 수 있다.

[0237] 도 3은 일 실시예에 따른 전자 장치의 구조도이다.

[0238] 전자 장치(300)는 전술된 부호화 장치(100) 및 복호화 장치(200)에 대응할 수 있다. 말하자면, 전자 장치(300)는 전술된 부호화 장치(100)의 부호화의 기능 및 복호화 장치(200)의 복호화의 기능을 수행할 수 있다.

[0239] 전자 장치(300)는 버스(390)를 통하여 서로 통신하는 처리부(310), 메모리(330), 사용자 인터페이스(User Interface; UI) 입력 디바이스(350), UI 출력 디바이스(360) 및 저장소(storage)(340)를 포함할 수 있다. 또한, 전자 장치(300)는 네트워크(399)에 연결되는 통신부(320)를 더 포함할 수 있다.

[0240] 처리부(310)는 중앙 처리 장치(Central Processing Unit; CPU), 메모리(330) 또는 저장소(340)에 저장된 프로세싱(processing) 명령어(instruction)들을 실행하는 반도체 장치일 수 있다. 처리부(310)는 적어도 하나의 하드웨어 프로세서일 수 있다.

[0241] 처리부(310)는 전자 장치(300)로 입력되거나, 전자 장치(300)에서 출력되거나, 전자 장치(300)의 내부에서 사용되는 신호, 데이터 또는 정보의 생성 및 처리를 수행할 수 있고, 신호, 데이터 또는 정보에 관련된 검사, 비교 및 판단 등을 수행할 수 있다. 말하자면, 실시예에서 데이터 또는 정보의 생성 및 처리와, 데이터 또는 정보에 관련된 검사, 비교 및 판단은 처리부(310)에 의해 수행될 수 있다.

[0242] 처리부(310)는 인터 예측부(110), 인트라 예측부(120), 스위치(115), 감산기(125), 변환부(130), 양자화부(140), 엔트로피 부호화부(150), 역양자화부(160), 역변환부(170), 가산기(175), 필터부(180) 및 참조 픽처 버퍼(190)를 포함할 수 있다.

[0243] 또한, 처리부(310)는 엔트로피 복호화부(210), 역양자화부(220), 역변환부(230), 인트라 예측부(240), 인터 예측부(250), 스위치(245), 가산기(255), 필터부(260) 및 참조 픽처 버퍼(270)를 포함할 수 있다.

[0244] 인터 예측부(110), 인트라 예측부(120), 스위치(115), 감산기(125), 변환부(130), 양자화부(140), 엔트로피 부호화부(150), 역양자화부(160), 역변환부(170), 가산기(175), 필터부(180) 및 참조 픽처 버퍼(190) 중 적어도 일부는 프로그램 모듈들일 수 있으며, 외부의 장치 또는 시스템과 통신할 수 있다. 프로그램 모듈들은 운영 체제, 응용 프로그램 모듈 및 기타 프로그램 모듈의 형태로 전자 장치(300)에 포함될 수 있다.

[0245] 또한, 엔트로피 복호화부(210), 역양자화부(220), 역변환부(230), 인트라 예측부(240), 인터 예측부(250), 스위치(245), 가산기(255), 필터부(260) 및 참조 픽처 버퍼(270) 중 적어도 일부는 프로그램 모듈들일 수 있으며, 외부의 장치 또는 시스템과 통신할 수 있다. 프로그램 모듈들은 운영 체제, 응용 프로그램 모듈 및 기타 프로그램 모듈의 형태로 전자 장치에 포함될 수 있다.

[0246] 프로그램 모듈들은 물리적으로는 여러 가지 공지의 기억 장치 상에 저장될 수 있다. 또한, 이러한 프로그램 모

듈 중 적어도 일부는 전자 장치(300)와 통신 가능한 원격 기억 장치에 저장될 수도 있다.

[0247] 프로그램 모듈들은 일 실시예에 따른 기능 또는 동작을 수행하거나, 일 실시예에 따른 추상 데이터 유형을 구현하는 루틴(routine), 서브루틴(subroutine), 프로그램, 오브젝트(object), 컴포넌트(component) 및 데이터 구조(data structure) 등을 포괄할 수 있지만, 이에 제한되지는 않는다.

[0248] 프로그램 모듈들은 전자 장치(300)의 적어도 하나의 프로세서(processor)에 의해 수행되는 명령어(instruction) 또는 코드(code)로 구성될 수 있다.

[0249] 처리부(310)는 인터 예측부(110), 인트라 예측부(120), 스위치(115), 감산기(125), 변환부(130), 양자화부(140), 엔트로피 부호화부(150), 역양자화부(160), 역변환부(170), 가산기(175), 필터부(180) 및 참조 픽처 버퍼(190)의 명령어 또는 코드를 실행할 수 있다.

[0250] 또한, 처리부(310)는 엔트로피 복호화부(210), 역양자화부(220), 역변환부(230), 인트라 예측부(240), 인터 예측부(250), 스위치(245), 가산기(255), 필터부(260) 및 참조 픽처 버퍼(270)의 명령어 또는 코드를 실행할 수 있다.

[0251] 저장부는 메모리(330) 및/또는 저장소(340)를 나타낼 수 있다. 메모리(330) 및 저장소(340)는 다양한 형태의 휘발성 또는 비휘발성 저장 매체일 수 있다. 예를 들면, 메모리(330)는 롬(ROM)(331) 및 램(RAM)(332) 중 적어도 하나를 포함할 수 있다.

[0252] 저장부는 전자 장치(300)의 동작을 위해 사용되는 데이터 또는 정보를 저장할 수 있다. 실시예에서, 전자 장치(300)가 갖는 데이터 또는 정보는 저장부 내에 저장될 수 있다.

[0253] 예를 들면, 저장부는 픽처, 블록, 리스트, 움직임 정보, 인터 예측 정보 및 비트스트림 등을 저장할 수 있다.

[0254] 전자 장치(300)는 컴퓨터에 의해 독출(read)될 수 있는 기록 매체를 포함하는 컴퓨터 시스템에서 구현될 수 있다.

[0255] 기록 매체는 전자 장치(300)가 동작하기 위해 요구되는 적어도 하나의 모듈을 저장할 수 있다. 메모리(330)는 적어도 하나의 모듈을 저장할 수 있고, 적어도 하나의 모듈이 처리부(310)에 의하여 실행되도록 구성될 수 있다.

[0256] 전자 장치(300)의 데이터 또는 정보의 통신과 관련된 기능은 통신부(320)를 통해 수행될 수 있다.

[0257] 예를 들면, 통신부(320)는 비트스트림을 다른 전자 장치로 전송할 수 있다.

[0258] 도 4는 일 예에 따른 풀 메시 토폴로지에 기반하는 파라미터의 공유를 나타낸다.

[0259] 도 4에서는, 파라미터를 공유하는 복수의 계산 노드들이 도시되었다. 계산 노드는 전술된 전자 장치(300)일 수 있다. 말하자면, 계산 노드는 분산 처리 시스템에서 분산 처리를 수행하는 전자 장치(300)를 의미할 수 있다.

[0260] 이하에서, 용어들 "파라미터", "학습 파라미터", "가중치(weight) 파라미터" 및 "커널(kernel) 파라미터"의 의미들은 동일 또는 유사할 수 있다. 또한, 상기의 용어들은 서로 교체되어 사용될 수 있다. 또한, 상기의 "파라미터"는 "가중치"를 나타낼 수도 있다.

[0261] 이하에서, 용어들 "학습(learning)" 및 "트레이닝(training)"의 의미들은 동일할 수 있다. 또한, 상기의 용어들은 서로 교체되어 사용될 수 있다.

[0262] 도 5는 일 예에 따른 스타 토폴로지에 기반하는 파라미터의 공유를 나타낸다.

[0263] 도 5에서는, 파라미터를 공유하는 복수의 계산 노드들이 도시되었고, 파라미터의 공유를 위해 공유 장소로서 사용되는 파라미터 서버가 도시되었다.

[0264] 딥 러닝 모델에 대한 분산 학습을 위해서는, 학습의 과정에서 딥 러닝 모델의 파라미터들이 다수의 계산 노드들 간에 공유되어야 할 수 있다.

[0265] 파라미터의 공유 방법으로서, 풀 메시(full mesh) 토폴로지(topology)에 기반하는 공유 방법과, 스타(star) 토폴로지에 기반하는 공유 방법이 사용될 수 있다.

[0266] 풀 메시 토폴로지에 기반하는 공유 방법이 사용될 경우, 하나의 계산 노드가 모든 다른 계산 노드들에게 직접적으로 파라미터를 전달할 수 있다.

[0267] 스타 토폴로지에 기반하는 공유 방법이 사용될 경우, 모든 계산 노드들은 공유 장소인 파라미터 서버를 이용하여 파라미터를 독출(read) 및 기입(write)할 수 있다.

[0268] 스타 토폴로지에 기반하는 공유 방법이 사용될 경우, 중앙의 파라미터 서버가 파라미터의 갱신 및 공유를 관리할 수 있다. 따라서, 파라미터의 공유를 위해 사용되는 저장소에서, 동시성(concurrency) 및 동기화(synchronization)에 대한 제어가 요구된다. 데이터에 대한 1 회의 학습 과정인 시대(epoch)가 반복(iterate)됨에 따라, 계산 노드들 간의 파라미터의 동기화가 요구될 수 있다.

[0269] 2 개의 공유 방법들의 모두에서, 토폴로지를 구성하는 계산 노드들의 개수가 증가함에 따라, 파라미터의 공유를 위한 계산 노드들 간의 통신의 횟수가 증가할 수 있다. 특히, 풀 메시 토폴로지에 기반하는 공유 방법이 사용될 경우, 파라미터의 공유를 위해 요구되는 통신의 횟수는 계산 노드들의 개수의 제곱에 비례할 수 있다.

[0270] 일반적으로, 딥 러닝 네트워크의 가중치 파라미터 셋 Θ에 대한 경사의 조절을 위해 경사 하강(gradient descent) 방식이 이용된다.

[0271] 경사 하강 방식에서, 에너지 함수 J(Θ)는 딥 러닝 네트워크의 예측 값 및 실제의 결과 값 간의 차이를 정의할 수 있다. 경사 하강 방식은 에너지 함수 J(Θ)를 최소화 하기 위해 기울기 ∇J(Θ) 를 이용하여 Θ를 학습하는 과정일 수 있다.

[0272] 예를 들면, 경사 하강 방식에서의 Θ의 학습은 아래의 수식 2와 같이 표현될 수 있다.

[0273] [수식 2]

[0274] $$\theta_{t+1} = \theta_t - \mu \nabla J(\theta_t) = \theta_t + \Delta\theta_t$$

[0275] Θ_t는 t 번째의 학습에서의 파라미터 셋일 수 있다. 충분히 큰 t에서 Θ_t가 최적 파라미터 셋 Θ_{opt}로 수렴한다는 것이 기대될 수 있다. μ는 스탭 크기일 수 있다.

[0276] 딥 러닝 모델의 파라미터 셋 내의 파라미터들의 개수는 수 십만을 상회할 수도 있다. 또한, 딥 러닝 모델의 학습을 위해 방대한 양의 입력 데이터가 사용될 수 있으며, 이러한 입력 데이터를 사용하는 학습을 위해 요구되는 반복(iteration)의 횟수 또한 수천 번을 상회할 수 있다. 이러한 파라미터들의 개수, 입력 데이터의 양 및 반복의 횟수에 따라서, 딥 러닝 모델의 학습을 위해서는 방대한 횟수의 수식 2의 연산들이 요구될 수 있다.

[0277] 대규모의 딥 러닝 모델의 분산 학습에 이용되는 현재까지의 학습 모델 및 파라미터 공유 방식에 따르면, 분산 처리 시스템 내의 분산 컴퓨터들의 개수가 증가하고, 경사 하강 알고리즘에 따라 커널 파라미터의 학습이 반복됨에 따라, 계산 노드들 간에서 공유되어야 하는 파라미터들의 개수가 급격히 늘어날 수 있다.

[0278] 이 때, 네트워크의 한정된 대역 폭(band width) 내에서 방대한 양의 파라미터가 공유됨에 따라 발생하는 네트워크의 지연 문제는 전체의 학습 과정을 느리게 할 수 있다. 또한, 이러한 방대한 양의 파라미터의 공유를 위해 요구되는 계산 노드의 저장 공간 또한 증가할 수 있다.

[0279] 실시예에서는, 공유되는 파라미터를 압축하고, 압축된 파라미터를 전송하는 효율적인 분산 학습 알고리즘이 제공될 수 있다.

[0280] 실시예에서는, 파라미터를 전송하는 분산 컴퓨터(말하자면, 부호화 장치)에서는 학습된 파라미터에 대한 잔차에 대해서만 예측 부호화를 수행하여 부호화된 잔차를 생성할 수 있다. 파라미터를 수신하는 분산 컴퓨터(말하자면, 복호화 장치)에서는 전달된 부호화된 잔차 및 이전의 파라미터를 이용하여 파라미터를 재구축할 수 있다.

[0281] 실시예는, 중앙의 파라미터 서버가 하나의 분산 컴퓨터에서 학습된 파라미터를 다수의 분산 컴퓨터들로 전송하는 경우에도 적용될 수 있다.

[0282] 도 6은 일 실시예에 따른 갱신된 파라미터에 대한 정보를 제공하는 방법의 흐름도이다.

[0283] 실시예에서, 전자 장치(300)는 부호화 장치(100)의 기능을 수행할 수 있다.

[0284] 분산 처리 시스템은 신경망의 학습을 분산하여 수행하는 복수의 전자 장치들을 포함할 수 있다. 아래에서, 전자 장치(300)는 복수의 전자 장치들 중 하나일 수 있다. 말하자면, 아래의 실시예는 분산 처리 시스템의 복수의 전자 장치들 중 하나의 전자 장치에서 수행되는 동작을 나타낼 수 있다.

[0285] 단계(610)에서, 분산 처리 시스템 내의 전자 장치(300)의 처리부(310)는 신경망의 학습을 수행할 수 있다.

[0286] 신경망은 딥 러닝을 위한 신경망일 수 있다.

[0287] 학습의 수행에 의해 신경망의 파라미터가 갱신될 수 있다.

[0288] 단계(620)에서, 전자 장치(300)의 처리부(310)는 신경망의 파라미터의 갱신을 위한 정보를 생성할 수 있다.

[0289] 신경망의 파라미터의 갱신을 위한 정보는 전자 장치(300)에서 이루어진 갱신된 파라미터를 다른 전자 장치들과 공유하기 위한 정보일 수 있다.

[0290] 처리부(310)는 학습의 수행에 의해 갱신된 파라미터의 상태 및 상기의 학습의 이전의 파라미터의 상태에 기반하여 파라미터의 갱신을 위한 정보를 생성할 수 있다.

[0291] 여기에서, 파라미터의 상태는 파라미터의 값을 의미할 수 있다. 말하자면, 처리부(310)는 학습의 수행에 의해 갱신된 파라미터의 값 및 상기의 학습의 이전의 파라미터의 값에 기반하여 파라미터의 갱신을 위한 정보를 생성할 수 있다.

[0292] 파라미터의 갱신을 위한 정보는 파라미터의 잔차(residual)를 포함할 수 있다. 파라미터의 잔차는 학습의 수행에 의해 갱신된 파라미터의 값 및 상기의 학습의 이전의 파라미터의 값 간의 차이일 수 있다. 처리부(310)는 파라미터의 잔차에 기반하여 파라미터의 갱신을 위한 정보를 생성할 수 있다.

[0293] 단계(630)에서, 전자 장치(300)의 통신부(320)는 파라미터의 갱신을 위한 정보를 분산 처리 시스템 내의 다른 장치로 전송할 수 있다.

[0294] 전자 장치(300)의 통신부(320)는 비트스트림을 분산 처리 시스템 내의 다른 장치로 전송할 수 있다. 비트스트림은 파라미터의 갱신을 위한 정보를 포함할 수 있다.

[0295] 다른 장치는 분산 처리 시스템 내에서 신경망의 학습을 수행하는 하나 이상의 다른 전자 장치들일 수 있다.

[0296] 다른 장치는 분산 처리 시스템 내의 파라미터 서버일 수 있다. 파라미터 서버는 파라미터의 갱신을 위한 정보를 분산 처리 시스템 내의 신경망의 학습을 수행하는 하나 이상의 다른 전자 장치들로 전달할 수 있다.

[0297] 도 7은 일 실시예에 따른 갱신된 파라미터에 대한 정보를 사용하여 파라미터의 갱신을 수행하는 방법의 흐름도이다.

[0298] 분산 처리 시스템은 신경망의 학습을 분산하여 수행하는 복수의 전자 장치들을 포함할 수 있다. 아래에서, 전자 장치(300)는 복수의 전자 장치들 중 하나일 수 있다. 말하자면, 아래의 실시예는 분산 처리 시스템의 복수의 전자 장치들 중 하나의 전자 장치에서 수행되는 동작을 나타낼 수 있다.

[0299] 실시예에서, 전자 장치(300)는 복호화 장치(200)의 기능을 수행할 수 있다.

[0300] 단계(710)에서, 전자 장치(300)의 통신부(320)는 분산 처리 시스템 내의 다른 장치로부터 파라미터의 갱신을 위한 정보를 수신할 수 있다.

[0301] 전자 장치(300)의 통신부(320)는 분산 처리 시스템 내의 다른 장치로부터 비트스트림을 수신할 수 있다. 비트스트림은 파라미터의 갱신을 위한 정보를 포함할 수 있다.

[0302] 다른 장치는 분산 처리 시스템 내에서 신경망의 학습을 수행하는 다른 전자 장치일 수 있다.

[0303] 다른 장치는 분산 처리 시스템 내의 파라미터 서버일 수 있다. 파라미터 서버는 분산 처리 시스템 내의 신경망의 학습을 수행하는 다른 전자 장치로부터 파라미터의 갱신을 위한 정보를 수신할 수 있다. 파라미터 서버는 파라미터의 갱신을 위한 정보를 전자 장치(300)로 전송할 수 있다.

[0304] 단계(720)에서, 처리부(310)는 파라미터의 갱신을 위한 정보를 사용하여 파라미터의 갱신을 수행할 수 있다.

[0305] 파라미터의 갱신을 위한 정보는, 다른 전자 장치에서 이루어진 학습의 수행에 의해 파라미터가 갱신되면, 상기의 파라미터의 갱신을 전자 장치(300)에도 공유하기 위한 정보일 수 있다.

[0306] 파라미터의 갱신을 위한 정보는 파라미터의 잔차(residual)를 포함할 수 있다. 파라미터의 잔차는 다른 전자 장치에서 이루어진 학습의 수행에 의해 갱신된 파라미터의 값 및 상기의 학습의 이전의 파라미터의 값 간의 차이일 수 있다.

[0307] 처리부(310)는 파라미터의 갱신을 위한 정보에 기반하여 파라미터의 잔차를 생성할 수 있다. 처리부(310)는 파라미터의 잔차를 파라미터에 더함으로써 파라미터를 갱신할 수 있다.

[0308] 도 8은 일 예에 따른 딥 러닝의 파라미터 셋의 갱신을 나타낸다.

[0309] 도 8에서는, 기존의 시간 t에서의 커널 파라미터 셋 및 커널 파라미터 잔차 신호 셋을 서로 합함으로써 시간 t+1에서의 커널 파라미터 셋이 생성되는 과정이 도시되었다. 말하자면, 커널 파라미터 잔차 신호 셋을 사용하여 특정된 시점에서의 커널 파라미터 셋이 다음 시점에서의 커널 파라미터 셋으로 갱신될 수 있다. 여기에서 커널 파라미터 잔차 신호 셋은 비트스트림을 통해 전달된 다른 전자 장치로부터의 정보일 수 있으며, 다른 전자 장치에서의 신경망의 학습을 통해 생성된 정보일 수 있다.

[0310] 도 6을 참조하여 전술된 실시예 및 도 7을 참조하여 전술된 실시예에서 갱신의 대상인 파라미터는 복수일 수 있다.

[0311] 복수의 파라미터들은 파라미터 셋을 구성할 수 있다. 도 7 및 도 8을 참조하여 설명된 파라미터의 갱신을 위한 정보는 복수의 파라미터들의 잔차 신호 셋일 수 있다. 또는, 파라미터의 갱신을 위한 정보는 복수의 파라미터들의 잔차 신호 셋을 포함할 수 있다. 또는, 파라미터의 갱신을 위한 정보는 복수의 파라미터들의 잔차 신호 셋에 기반하여 생성될 수 있고, 반대로 파라미터의 갱신을 위한 정보에 기반하여 복수의 파라미터들의 잔차 신호 셋이 생성될 수 있다.

[0312] 여기에서, 복수의 파라미터들은 딥 러닝 모델의 하나의 계층을 구성하는 딥 러닝 파라미터들일 수 있다.

[0313] 전자 장치(300)는 비트스트림에 대한 복호화를 수행함으로써 복수의 파라미터들의 잔차 신호 셋을 획득할 수 있다.

[0314] 잔차 신호는 수식 2에서의 $\Delta\Theta_t$에 해당할 수 있다.

[0315] Θ_{t+1}는 학습이 반복(iteration)될 때마다 Θ_t로부터 갱신될 수 있다.

[0316] 복수의 파라미터들의 잔차들은 압축될 수 있다. 말하자면, 파라미터의 갱신을 위한 정보는 복수의 파라미터들의 압축된 잔차들을 포함할 수 있다. Θ_{t+1}의 갱신의 크기는 크지 않기 때문에 압축된 Θ_{t+1} 보다는 압축된 Θ_t를 전송함에 따라 전송되는 데이터의 크기가 감소될 수 있다.

[0317] 예를 들면, 단계(620)에서, 전자 장치(300)의 처리부(310)는 복수의 파라미터들의 잔차들에 부호화를 수행함으로써 복수의 파라미터들의 압축된 잔차들을 생성할 수 있고, 복수의 파라미터들의 압축된 잔차들을 포함하는 파라미터의 갱신을 위한 정보를 생성할 수 있다.

[0318] 전자 장치(300)의 처리부(310)는 사전에 가지고 있는 Θ_t에 복호화된 $\Delta\Theta_t$을 더함으로써 Θ_{t+1}를 재구축할 수 있다.

[0319] 예를 들면, 단계(720)에서, 처리부(310)는 파라미터의 갱신을 위한 정보에 포함된 복수의 파라미터들의 압축된 잔차들에 대한 복호화를 수행함으로써 복수의 파라미터들의 잔차들을 획득할 수 있고, 복수의 파라미터들의 잔차들을 사용하여 복수의 파라미터들의 갱신을 수행할 수 있다.

[0320] 복수의 파라미터들의 잔차들에 대한 부호화 및 복호화에 있어서, 전술된 영상의 블록에 대한 부호화 방법 및 복호화 방법이 사용될 수 있다. 말하자면, 복수의 파라미터들은 대상 블록으로 간주될 수 있다. 복수의 파라미터들의 값들은 대상 블록의 샘플들로 간주될 수 있고, 복수의 파라미터들의 잔차들은 대상 블록에 대한 잔차 블록으로 간주될 수 있다.

[0321] CNN의 컨벌루션 레이어에서는 NxN 커널의 블록의 파라미터들이 학습될 수 있다. 학습된 커널 파라미터들에 대한 필터링이 적용될 수 있으며, 필터링이 적용된 커널 파라미터들은 다음의 레이어의 특징 맵(feature map)을 생성할 수 있다.

[0322] 일반적으로, CNN에서는 하나의 커널 파라미터에 대한 학습이 수행되지는 않으며, 각 계층에서 다수의 커널 파라미터들에 대한 학습이 이루어질 수 있다. 이러한 다수의 커널 파라미터들에 대한 학습을 통해 파라미터 셋이 구성될 수 있다.

[0323] 도 8에서는, CNN의 하나의 레이어를 구성하는 24 개의 커널 파라미터들이 24 개의 블록들로서 표시되었다. 말하자면, 블록은 커널 파라미터에 대응할 수 있고, 커널 파라미터는 블록에 대응할 수 있다. 블록의 크기는 NxN일

수 있다. 블록들의 크기는 동일할 수 있다. 24 개의 블록들은 영상을 구성할 수 있다. 실시예의 과정들을 통해, 충분한 반복들의 이후에, 커널 파라미터 셋은 보다 적은 노이즈를 갖는 형태로 수렴될 수 있다.

[0324] 전술된 것과 같이, 파라미터 셋은 복수의 파라미터들을 포함할 수 있고, 복수의 파라미터들의 각 파라미터는 블록에 대응할 수 있다. 말하자면, 파라미터 셋은 복수의 블록들을 포함할 수 있다. 또한, 복수의 블록들의 크기들은 동일할 수 있다.

[0325] 파라미터 셋은 복수의 블록들을 포함하는 대상 영상으로 간주될 수 있다. 말하자면, 블록은 복수일 수 있고, 복수의 블록들은 대상 영상을 구성할 수 있다. 이전의 실시예들에서 전술된 대상 영상에 대한 처리 방법이 본 실시예에도 적용될 수 있다. 여기에서, 처리 방법은 부호화 및 복호화를 포함할 수 있다.

[0326] 예를 들면, 단계(720)에서, 처리부(310)는 파라미터의 갱신을 위한 정보에 포함된 복수의 파라미터들의 압축된 잔차들에 대한 복호화를 수행함으로써 복수의 파라미터들의 잔차들을 획득할 수 있고, 복수의 파라미터들의 잔차들을 사용하여 복수의 파라미터들의 갱신을 수행할 수 있다. 이 때, 복수의 블록들은 대상 영상을 구성할 수 있고, 복호화에 있어서 전술된 실시예들에서의 대상 영상에 대한 복호화 방법이 사용될 수 있다.

[0327] 대상 영상 내의 블록은 이전의 실시예들에서 전술된 대상 블록으로 간주될 수 있다. 이전의 실시예들에서 전술된 대상 블록에 대한 처리 방법이 본 실시예의 블록에도 적용될 수 있다. 여기에서, 처리 방법은 부호화 및 복호화를 포함할 수 있다.

[0328] 여기에서, 전자 장치(300)는 파라미터의 부호화에 있어서, 블록 또는 신호에 대한 변환, 양자화, 스캐닝 및 엔트로피 부호화 중 하나 이상을 사용할 수 있다. 또한, 전자 장치(300)는 파라미터의 복호화에 있어서, 블록 또는 신호에 대한 엔트로피 복호화, 스캐닝, 역양자화 및 역변환 중 하나 이상을 사용할 수 있다. 이러한 부호화 및 복호화에 있어서 적어도 일부의 과정은 변경 및/또는 생략될 수 있다.

[0329] 도 9는 일 예에 따른 양자화, 레스터 스캔 및 엔트로피 코딩의 과정을 나타낸다.

[0330] 블록은 특정된 크기를 갖는 필터의 경사들일 수 있다.

[0331] 도 9에서, 4-by-4(또는, 4x4) 필터의 경사(gradient)들에 대한 양자화의 결과인 양자화된 경사들이 도시되었고, 상기의 양자화된 경사들에 대한 역양자화의 결과인 역양자화된 경사들이 도시되었다. 여기에서, 필터는 커널을 의미할 수 있다. 경사는 잔차 신호, 즉 $\Delta\Theta_t$을 의미할 수 있다.

[0332] 양자화된 경사는 양자화 인덱스로도 명명될 수 있다.

[0333] 또한, 도 9에서는, 상기의 양자화된 경사들에 대하여, 2차원(Dimension; D)로부터 1D로의 변환을 위한 스캐닝(즉, 래스터 스캔(raster scan)이 적용되었고, 스캐닝에 의해 1D로 스캐닝된 양자화된 경사들이 생성되었다. 1D로 스캐닝된 양자화된 경사들에 대해 엔트로피 코딩이 적용됨으로써 엔트로피 코딩된 양자화된 경사들이 생성될 수 있다.

[0334] 예를 들면, 파라미터의 갱신을 위한 정보는 엔트로피 코딩된 양자화된 경사를 포함할 수 있다.

[0335] 예를 들면, 단계(620)에서, 전자 장치(300)의 처리부(310)는 필터의 경사들에 대한 양자화를 수행함으로써 양자화된 경사들을 생생할 수 있다. 처리부(310)는 양자화된 경사들에 대한 2D로부터 1D로의 스캐닝을 수행함으로써 스캐닝된 정보를 생성할 수 있다. 스캐닝된 정보는 스캐닝된 양자화된 경사들을 포함할 수 있다. 처리부(310)는 스캐닝된 정보에 대한 엔트로피 부호화를 수행함으로써 파라미터의 갱신을 위한 정보를 생성할 수 있다. 파라미터의 갱신을 위한 정보는 엔트로피 부호화된 양자화된 경사들을 나타낼 수 있다.

[0336] 예를 들면, 단계(720)에서, 전자 장치(300)의 처리부(310)는 비트스트림으로부터 파라미터의 갱신을 위한 정보를 획득할 수 있다. 처리부(310)는 파라미터의 갱신을 위한 정보에 대한 엔트로피 복호화를 수행함으로써 스캐닝된 정보를 생성할 수 있다. 말하자면, 처리부(310)는 파라미터의 갱신을 위한 정보에 기반하여 스캐닝된 정보를 생성할 수 있다. 스캐닝된 정보는 스캐닝된 양자화된 경사들을 포함할 수 있다. 처리부(310)는 스캐닝된 정보에 대한 1D로부터 2D로의 스캐닝을 수행함으로써 양자화된 경사들을 생성할 수 있다. 처리부(310)는 양자화된 경사들에 대한 역양자화를 수행함으로써 재구축된 경사들을 생성할 수 있다. 재구축된 경사들은 복수의 파라미터들의 잔차들일 수 있다.

[0337] 도 10은 일 예에 따른 제로 서브 블록 및 제로 레이어를 사용하는 엔트로피 부호화를 나타낸다.

[0338] 신경망의 학습의 반복이 수행됨에 따라, 수식 2가 나타내는 학습의 결과가 수렴할 수 있다. 수렴에 의해 $\Delta\Theta_t$가

충분히 작아짐에 따라, 현재의 반복에서 블록 내의 양자화 인덱스들(즉, 양자화된 경사들)의 0의 분포가 변할 수 있다. 예를 들면, 신경망의 학습의 반복이 수행됨에 따라 블록 내에서 0의 값을 갖는 양자화된 경사들의 개수가 증가할 수 있다.

[0339] 이러한 증가에 따라, 블록 내의 특정된 영역의 양자화 인덱스들의 값들이 모두 0이 될 수 있다. 또는, 블록 내의 특정된 영역의 양자화 인덱스들의 값들이 공유될 필요가 없을 수 있다. 특정된 영역의 양자화 인덱스들의 값들이 모두 0이거나 특정된 영역의 양자화 인덱스들의 값들이 공유될 필요가 없으면, 특정된 영역은 0의 값을 갖는 제로 서브 블록으로 간주될 수 있다. 스캐닝된 정보는 특정된 영역에 대하여 제로 서브 블록을 포함할 수 있다. 도 10에서, 스캐닝된 정보 내의 제로 서브 블록은 "(제로 서브 블록)"으로 도시되었다.

[0340] 예를 들면, 단계(620)에서, 전자 장치(300)의 처리부(310)는 특정된 영역의 서브 블록 내의 양자화된 경사들의 값들이 모두 0이면, 상기의 0의 값들을 갖는 양자화된 경사들 대신 제로 서브 블록을 생성할 수 있다. 제로 서브 블록은 특정된 서브 블록의 0 의 값들을 갖는 양자화된 경사들 대신 스캐닝된 정보 내에 포함될 수 있다. 여기에서, 제로 서브 블록은 블록 내의 특정된 영역을 나타낼 수 있고, 상기의 특정된 영역 내의 양자화된 경사들의 값들이 모두 0임을 나타낼 수 있다.

[0341] 예를 들면, 단계(720)에서, 전자 장치(300)의 처리부(310)는 스캐닝된 정보 내에 제로 서브 블록이 존재할 경우, 제로 서브 블록이 나타내는 특정된 영역 내의 양자화된 경사들을 0의 값으로 채울 수 있다.

[0342] 또한, 이러한 증가에 따라, 블록의 양자화 인덱스들의 값들이 모두 0이 될 수 있다. 또는, 블록의 양자화 인덱스들의 값들이 공유될 필요가 없을 수 있다. 블록의 양자화 인덱스들의 값들이 모두 0이거나, 블록의 양자화 인덱스들의 값들이 공유될 필요가 없으면, 블록은 0의 값을 갖는 제로 블록으로 간주될 수 있다. 이러한 경우, 블록에 대한 스캐닝된 양자화된 경사들 대신 제로 블록이 사용될 수 있다.

[0343] 예를 들면, 단계(620)에서, 전자 장치(300)의 처리부(310)는 블록의 양자화된 경사들의 값들이 모두 0이거나, 블록의 양자화된 경사들의 값들이 공유될 필요가 없으면, 0의 값들을 갖는 스캐닝된 양자화된 경사들 대신 제로 블록을 생성할 수 있다. 스캐닝된 정보는 제로 블록을 포함할 수 있다. 제로 블록은 블록 내의 양자화된 경사들의 값들이 모두 0임을 나타낼 수 있다.

[0344] 예를 들면, 단계(720)에서, 전자 장치(300)의 처리부(310)는 블록에 대한 스캐닝된 정보가 제로 블록일 경우, 블록 내의 모든 양자화된 경사들을 0의 값으로 채울 수 있다.

[0345] 또한, 이러한 증가에 따라, 레이어의 파라미터들의 값들이 모두 0이 될 수 있다. 또는, 레이어의 파라미터들의 값들이 공유될 필요가 없을 수 있다. 특정된 파라미터들의 값들이 모두 0이거나, 레이어의 파라미터들의 값들이 공유될 필요가 없으면, 레이어에 대한 양자화된 경사들 대신 제로 레이어가 전송될 수 있다. 제로 레이어는 레이어의 블록들이 모두 제로 블록임을 나타낼 수 있다. 말하자면, 제로 레이어는 레이어의 모든 블록들의 모든 양자화된 경사들의 값들이 0임을 나타낼 수 있다. 도 10에서, 스캐닝된 정보 내의 제로 레이어는 "(제로 레이어)"로 도시되었다.

[0346] 예를 들면, 단계(620)에서, 전자 장치(300)의 처리부(310)는 레이어의 파라미터의 값들이 모두 0이거나, 레이어의 파라미터의 값들이 공유될 필요가 없으면, 레이어의 블록들의 스캐닝된 양자화된 경사들 대신 제로 레이어를 생성할 수 있다. 스캐닝된 정보는 제로 레이어를 포함할 수 있다. 제로 레이어는 레이어의 블록들의 양자화된 경사들의 값들이 모두 0임을 나타낼 수 있다.

[0347] 예를 들면, 단계(720)에서, 전자 장치(300)의 처리부(310)는 레이어에 대한 스캐닝된 정보가 제로 레이어인 경우, 레이어의 모든 블록들의 모든 양자화된 경사들을 0의 값으로 채울 수 있다.

[0348] 전술된 것과 같은 특정된 정보의 공유가 필요한지 여부는 계산 노드들 간의 네트워크의 성능에 기반하여 판단될 수 있다.

[0349] 복호화 장치(200)의 기능을 수행하는 전자 장치(300)는 파라미터에 대한 정보의 전송에 대한 네트워크의 성능을 예측할 수 있고, 예측된 성능에 기반하여 파라미터의 공유를 위한 전송의 여부를 결정할 수 있다. 여기에서, 파라미터는 서브 블록의 양자화된 경사, 블록의 양자화된 경사 및/또는 레이어의 양자화된 경사를 의미할 수 있다.

[0350] 예를 들면, 네트워크가 영상의 분류를 위해 사용되는 경우, 전자 장치(300)는 아래의 수식 3의 비용 함수 J가 최소화되도록 파라미터의 전송 여부를 결정할 수 있다.

[0351] [수식 3]

[0352] $J = C + \lambda B$

[0353] C는 네트워크의 분류 성능일 수 있다. B는 파라미터의 전송을 위해 요구되는 비트의 개수일 수 있다. λ는 상수일 수 있다.

[0354] 전자 장치(300)의 처리부(310)는 비용 함수 J가 최소가 되도록 파라미터의 전송 여부를 결정할 수 있다. 또는, 처리부(310)는 비용 함수가 J가 최소가 되도록 스캐닝된 정보 내에 제로 서브 블록, 제로 블록 또는 제로 레이어를 포함시킬 수 있다.

[0355] 도 11은 일 예에 따른 모델 병렬 처리 방식에서의 파라미터의 공유를 나타낸다.

[0356] 모델 병렬 처리 방식에서는 하나의 계산 노드가 대용량의 딥 러닝 모델을 처리하지 못할 수 있다. 모델 병렬 처리 방식에서는, 전자 장치(300)는 딥 러닝 모델의 일 부분에 대하여 연산된 파라미터를 반복(iteration)마다 다른 전자 장치와 공유해야 할 수 있다.

[0357] 신경망의 학습의 각 반복마다, 전자 장치(300)는 딥 러닝 모델의 일 부분에 대한 학습을 통해 상기의 일 부분에 대한 파라미터를 갱신할 수 있고, 갱신된 파라미터를 다른 전자 장치와 공유할 수 있다. 공유를 위해, 전자 장치(300)는 전술된 파라미터의 갱신을 위한 정보를 다른 장치들로 전송할 수 있다.

[0358] 또한, 다른 전자 장치에서도 딥 러닝 모델의 다른 부분에 대한 학습을 통해 상기의 다른 부분에 대한 다른 파라미터를 갱신할 수 있고, 갱신된 다른 파라미터를 전자 장치(300)와 공유할 수 있다. 공유를 위해, 다른 전자 장치는 다른 파라미터의 갱신을 위한 정보를 전자 장치(300)로 전송할 수 있다.

[0359] 따라서, 신경망의 학습의 각 반복마다, 복수의 전자 장치들에서 분할되어 생성된 갱신된 파라미터들이 복수의 전자 장치들 간에 공유될 수 있다.

[0360] 도 11에서는, 3 개의 계산 노드들(즉, 3 개의 전자 장치들)이 딥 러닝 모델의 3 개의 레이어들에 대한 학습을 분할하여 수행하는 것이 도시되었다. 즉, 계산 노드 1, 계산 노드 2 및 계산 노드 3이 레이어 1에 대한 학습, 레이어 2에 대한 학습 및 레이어 3에 대한 학습을 각각 담당할 수 있으며, 시간이 경과함에 따라 레이어에 대한 학습을 수행하면서, 레이어의 파라미터를 갱신할 수 있다.

[0361] 전자 장치(300)의 처리부(310)는 딥 러닝 모델의 복수의 레이어들 중 하나의 레이어에 대한 학습을 수행할 수 있다. 처리부(310)는 하나의 레이어에 대한 파라미터를 갱신할 수 있다. 처리부(310)는 학습에 의해 갱신된 파라미터의 값 및 상기의 학습의 이전의 파라미터의 값에 기반하여 레이어의 파라미터의 갱신을 위한 정보를 생성할 수 있다. 레이어의 파라미터의 갱신을 위한 정보는 다른 레이어에 대한 학습을 수행하는 다른 전자 장치들로 전송될 수 있다.

[0362] 이러한 파라미터의 갱신을 위한 정보의 전송은 학습의 각 반복마다 수행될 수 있다.

[0363] 전술된 것과 같이, 파라미터의 갱신을 위한 정보는 파라미터의 잔차를 포함하거나, 상기의 잔차를 나타낼 수 있다. 파라미터의 잔차는 현재의 반복에서의 파라미터의 값 및 이전의 반복에서의 파라미터의 값 간의 차이일 수 있다. 또는, 파라미터의 갱신을 위한 정보는 커널 필터의 잔차들을 포함할 수 있다.

[0364] 도 12는 일 예에 따른 데이터 병렬 처리 방식에서의 파라미터의 공유를 나타낸다.

[0365] 데이터 병렬 처리 방식에서는 학습을 위해 사용되는 입력 데이터가 분할(partition)될 수 있다. 계산 노드는 분할된 입력 데이터를 사용하여 딥 러닝 모델의 학습을 수행할 수 있다.

[0366] 말하자면, 복수의 전자 장치(300)들은 입력 데이터의 분할에 의해 생성된 분할된 입력 데이터를 각각 사용하여 딥 러닝 모델의 학습을 수행할 수 있다.

[0367] 스타 토폴로지 공유 방식이 사용될 때, 전자 장치(300)는 파라미터의 갱신을 위한 정보를 파라미터 서버로 전송할 수 있다. 파라미터 서버는 전송된 파라미터의 갱신을 위한 정보를 저장함으로써 전자 장치들에 의해 이루어진 학습을 기록할 수 있다. 파라미터 서버는 수신된 파라미터의 갱신을 위한 정보를 다른 전자 장치들로 전송할 수 있다. 다른 전자 장치는 수신된 파라미터의 갱신을 위한 정보를 사용하여 파라미터를 갱신할 수 있다. 이러한 전송 및 갱신을 통해 학습이 계속적으로 이루어질 수 있다.

[0368] 전자 장치(300)는 갱신된 파라미터를 파라미터 공유 서버로 전송할 수도 있다. 파라미터 공유 서버는 갱신된 파

라미터 및 저장된 파라미터를 사용하여 파라미터의 잔차만을 다른 전자 장치로 전송할 수 있다. 여기에서, 파라미터의 잔차는 파라미터 공유 서버 내에 저장된 파라미터의 값 및 전자 장치(300)로부터 전송된 갱신된 파라미터의 값 간의 차이일 수 있다.

[0369] 상술한 실시예들에서, 방법들은 일련의 단계 또는 유닛으로서 순서도를 기초로 설명되고 있으나, 본 발명은 단계들의 순서에 한정되는 것은 아니며, 어떤 단계는 상술한 바와 다른 단계와 다른 순서로 또는 동시에 발생할 수 있다. 또한, 당해 기술 분야에서 통상의 지식을 가진 자라면 순서도에 나타난 단계들이 배타적이지 않고, 다른 단계가 포함되거나, 순서도의 하나 또는 그 이상의 단계가 본 발명의 범위에 영향을 미치지 않고 삭제될 수 있음을 이해할 수 있을 것이다.

[0370] 이상 설명된 본 발명에 따른 실시예들은 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0371] 컴퓨터 판독 가능한 기록 매체는 본 발명에 따른 실시예들에서 사용되는 정보를 포함할 수 있다. 예를 들면, 컴퓨터 판독 가능한 기록 매체는 비트스트림을 포함할 수 있고, 비트스트림은 본 발명에 따른 실시예들에서 설명된 정보를 포함할 수 있다.

[0372] 컴퓨터 판독 가능한 기록 매체는 비-일시적 컴퓨터 판독 가능한 매체(non-transitory computer-readable medium)를 포함할 수 있다.

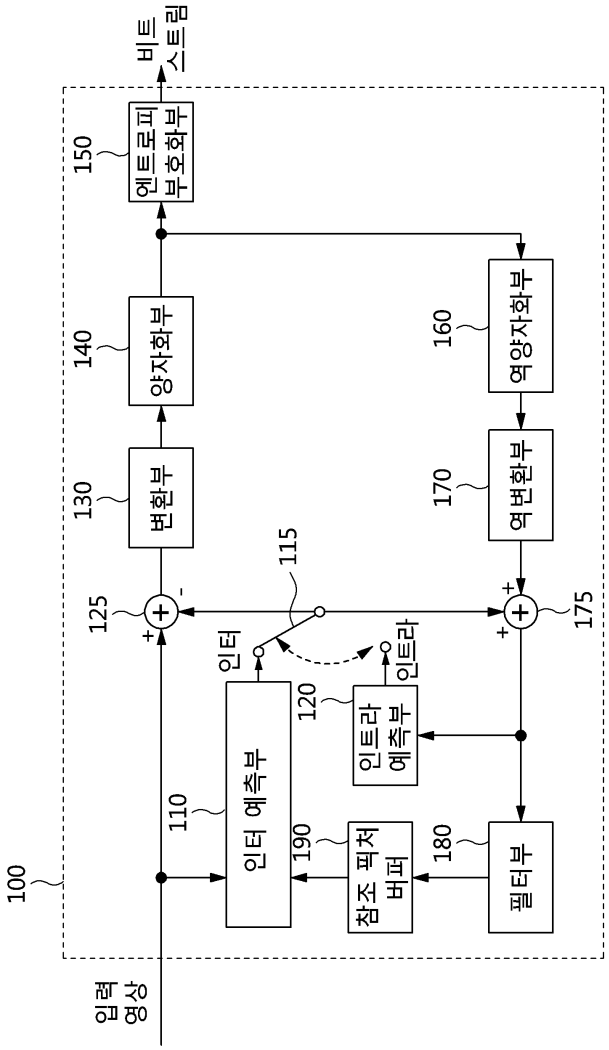
[0373] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0374] 이상에서 본 발명이 구체적인 구성요소 등과 같은 특정 사항들과 한정된 실시예 및 도면에 의해 설명되었으나, 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명이 상기 실시예들에 한정되는 것은 아니며, 본 발명이 속하는 기술분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형을 꾀할 수 있다.

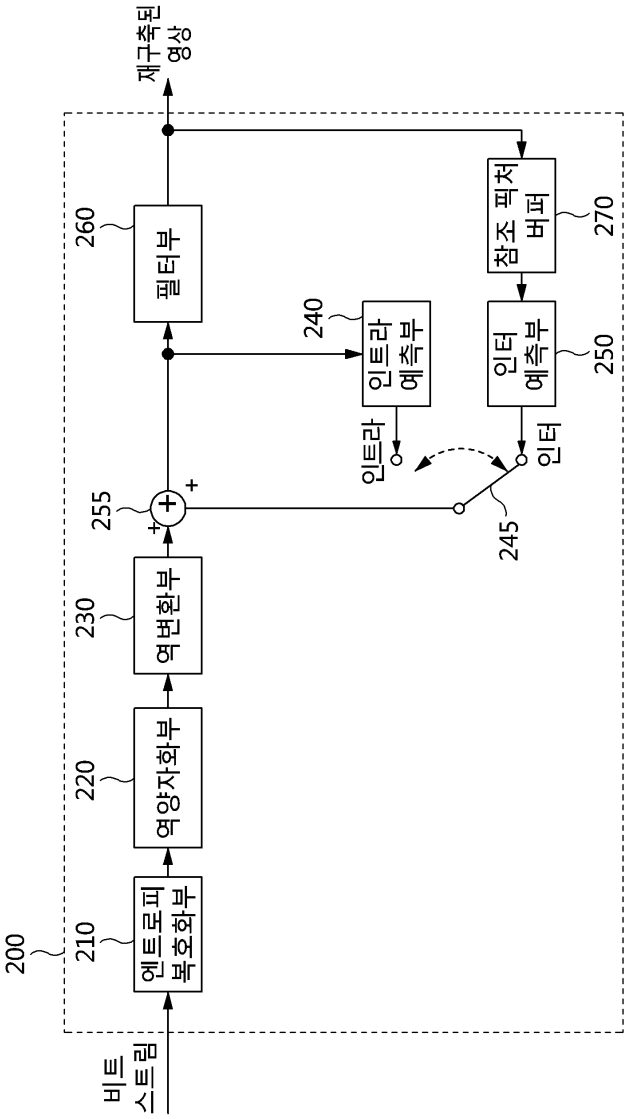
[0375] 따라서, 본 발명의 사상은 상기 설명된 실시예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐만 아니라 이 특허청구범위와 균등하게 또는 등가적으로 변형된 모든 것들은 본 발명의 사상의 범주에 속한다고 할 것이다.

도면

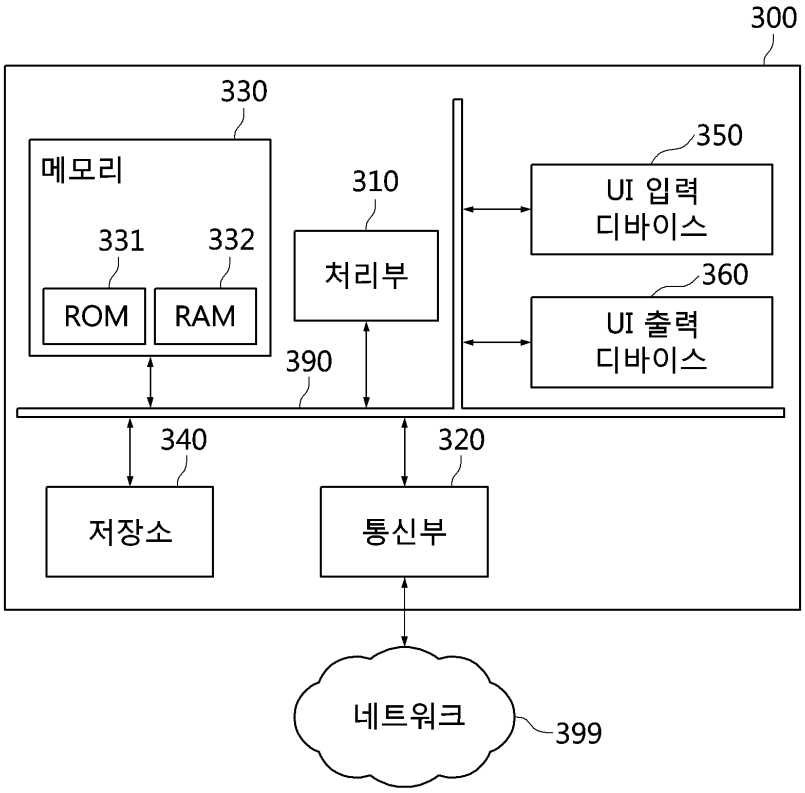
도면1


비트 스트림
엔트로피 부호화부
150
양자화부
140
변환부
130
125
115
인터
인트라
120
인터 예측부
110
인트라 예측부
190
참조 픽처 버퍼
180
필터부
역양자화부
160
역변환부
170
175
100
입력 영상

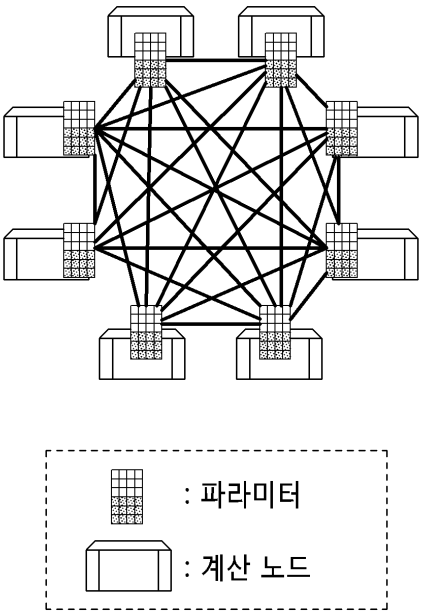
도면2


비트 스트림
200
엔트로피 복호화부
210
역양자화부
220
역변환부
230
+
+
255
245
인트라
인터
인트라 예측부
240
인터 예측부
250
참조 픽처 버퍼
270
필터부
260
재구축된 영상

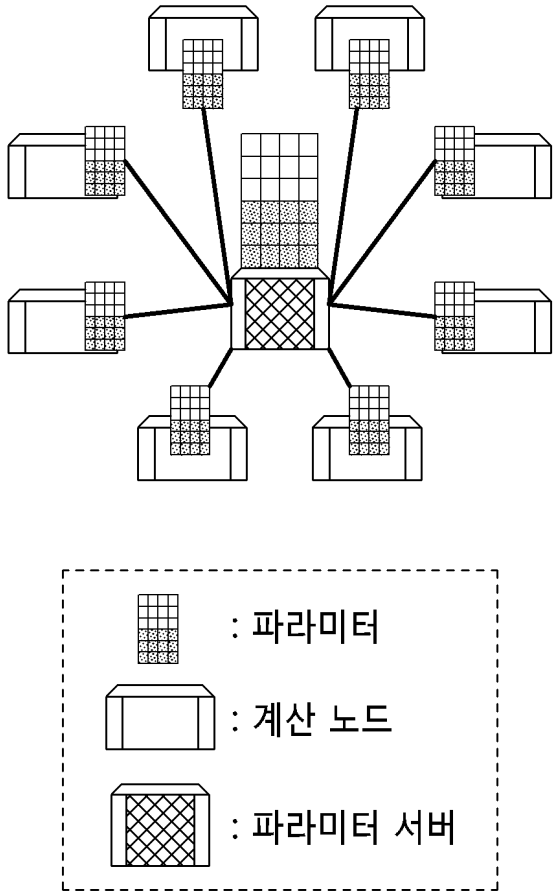
도면3


300
330
메모리
331
ROM
332
RAM
310
처리부
390
350
UI 입력
디바이스
360
UI 출력
디바이스
340
저장소
320
통신부
네트워크
399

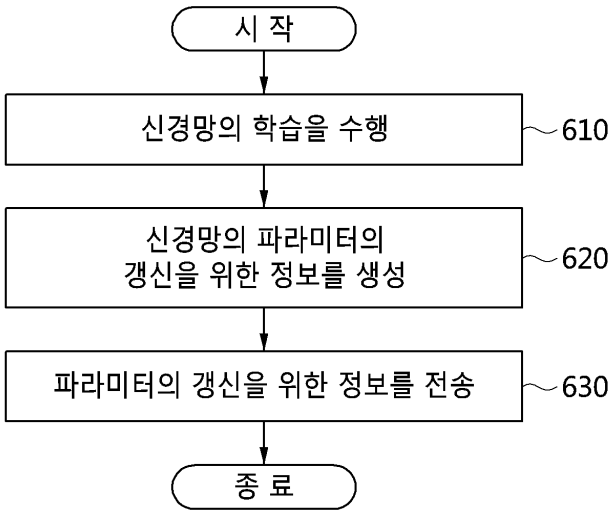
도면4


: 파라미터
: 계산 노드

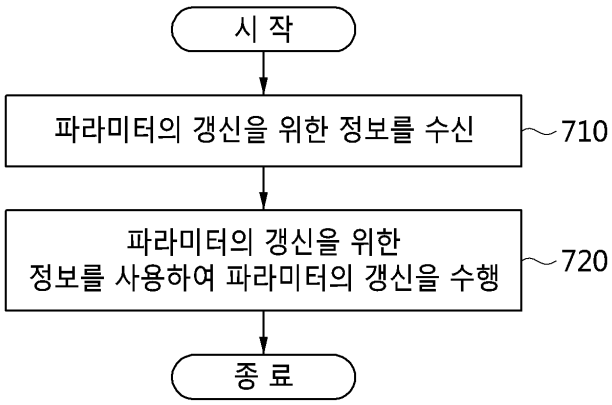
도면5


: 파라미터
: 계산 노드
: 파라미터 서버

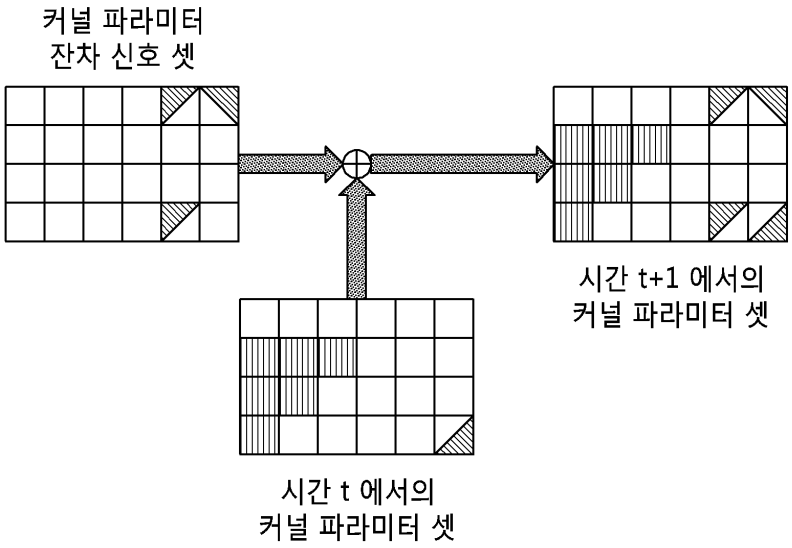
도면6


시 작
신경망의 학습을 수행
610
신경망의 파라미터의
갱신을 위한 정보를 생성
620
파라미터의 갱신을 위한 정보를 전송
630
종 료

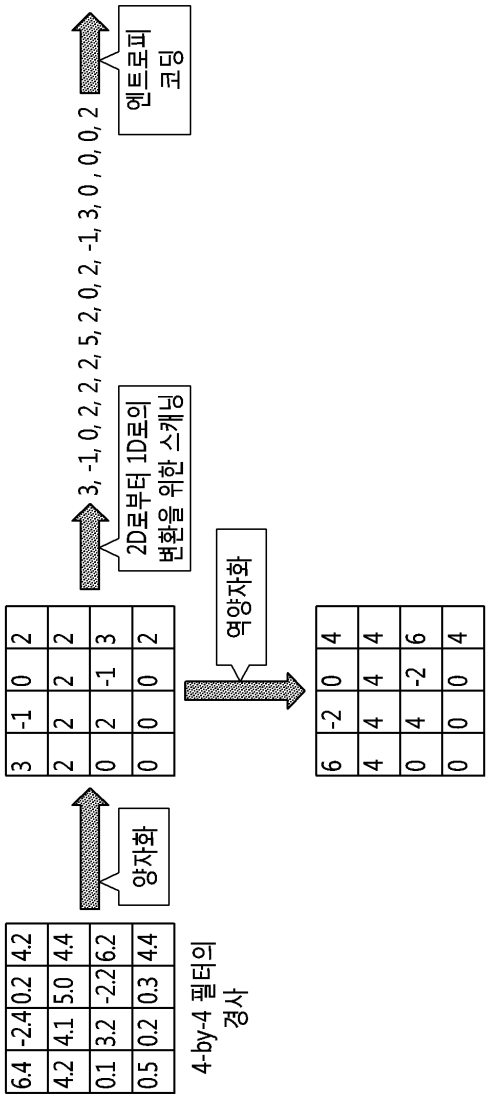
도면7


시 작
파라미터의 갱신을 위한 정보를 수신
710
파라미터의 갱신을 위한
정보를 사용하여 파라미터의 갱신을 수행
720
종 료

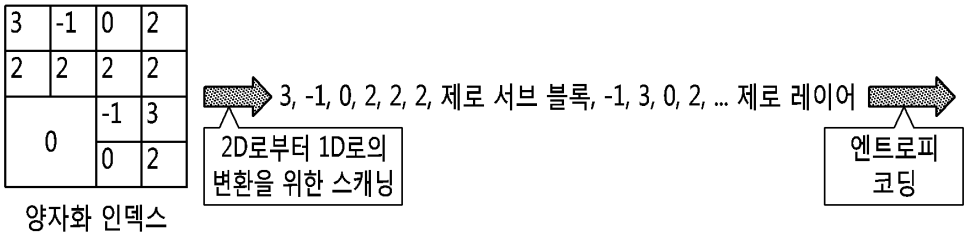
도면8


커널 파라미터
잔차 신호 셋
시간 t+1 에서의
커널 파라미터 셋
시간 t 에서의
커널 파라미터 셋

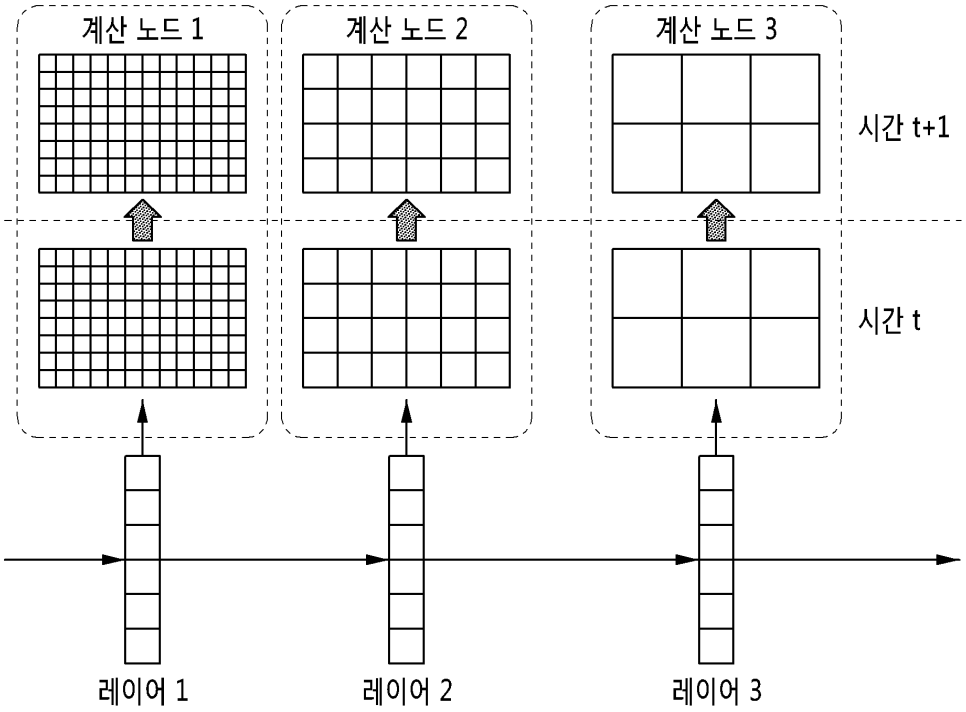
도면9


엔트로피 코딩
3, -1, 0, 2, 2, 2, 2, 5, 2, 0, 2, -1, 3, 0 , 0, 0, 2
2D로부터 1D로의 변환을 위한 스캐닝
역양자화
양자화
4-by-4 필터의 경사

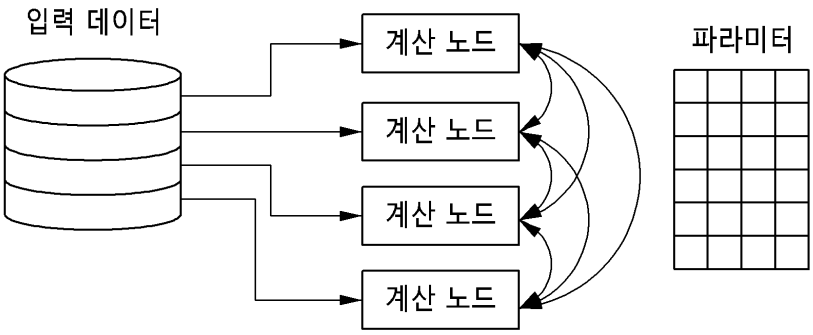
도면10


3, -1, 0, 2, 2, 2, 제로 서브 블록, -1, 3, 0, 2, ... 제로 레이어
2D로부터 1D로의 변환을 위한 스캐닝
엔트로피 코딩
양자화 인덱스

도면11


계산 노드 1
계산 노드 2
계산 노드 3
시간 t+1
시간 t
레이어 1
레이어 2
레이어 3

도면12


입력 데이터
계산 노드
계산 노드
계산 노드
계산 노드
파라미터