(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0178168
(43) 공개일자 2024년12월30일

(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) *G06N 3/0464* (2023.01)
G06N 3/048 (2023.01) *G06T 3/40* (2024.01)
G06V 10/46 (2022.01) *G06V 10/75* (2022.01)
G06V 10/762 (2022.01) *G06V 10/82* (2022.01)
(52) CPC특허분류
G06N 3/0495 (2023.01)
G06N 3/0464 (2023.01)
(21) 출원번호 10-2023-0102284
(22) 출원일자 2023년08월04일
심사청구일자 없음
(30) 우선권주장
1020230078358 2023년06월19일 대한민국(KR)

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
정승원
서울특별시 서초구 서초대로65길 13-10, 113동 1603호 (서초동,서초래미안아파트)
이홍재
서울특별시 동대문구 약령서길 52, 로망스빌 402호 (제기동)
유준상
경기도 성남시 분당구 서판교로58번길 11 (판교동)
(74) 대리인
리앤목특허법인

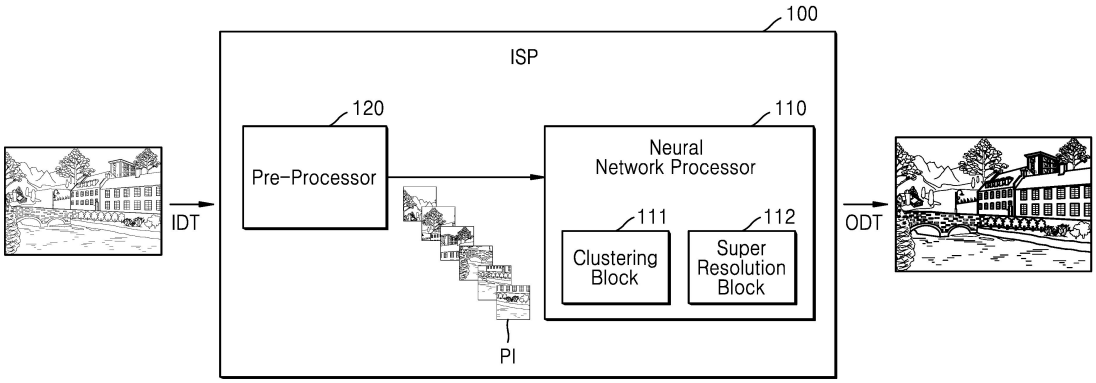
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **뉴럴 네트워크 프로세서를 포함하는 이미지 처리 장치 및 이미지 처리 방법**

(57) 요 약

본 개시의 기술적 사상에 따른 이미지 처리 장치는, 복수의 패치 이미지들 중 클러스터의 수에 따라 상기 복수의 패치 이미지들을 클러스터링하고, 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조(reference) 패치 이미지 및 쿼리(query) 패치 이미지로 구별하고, 상기 구별에 따라, 참조 패치 이미지 및 쿼리 패치 이미지에 대해 서로 상이한 비트수로 양자화를 수행할 수 있다.

대 표 도



(52) CPC특허분류
G06N 3/048 (2023.01)
G06T 3/4053 (2024.01)
G06V 10/462 (2023.08)
G06V 10/467 (2023.08)
G06V 10/751 (2023.08)
G06V 10/762 (2023.08)
G06V 10/82 (2022.01)
G06T 2207/20084 (2013.01)

명 세 서

청구범위

청구항 1

이미지 처리 장치에 있어서,

입력 이미지를 복수의 패치(patch) 이미지들로 분할하는 프리-프로세서; 및

상기 복수의 패치 이미지들 각각에 대해 이미지 처리 동작을 수행하여 출력 이미지를 생성하는 뉴럴 네트워크 프로세서;를 포함하고,

상기 뉴럴 네트워크 프로세서는,

상기 복수의 패치 이미지들 중 클러스터의 수에 따라 상기 복수의 패치 이미지들을 클러스터링하고, 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조(reference) 패치 이미지 및 쿼리(query) 패치 이미지로 구별하는 클러스터링 블록; 및

상기 각 클러스터의 상기 참조 패치 이미지인지 및 상기 쿼리 패치 이미지인지 여부에 따라, 상기 참조 패치 이미지 및 상기 쿼리 패치 이미지에 대해 서로 상이한 비트(bit) 수로 양자화를 수행하는 초해상화 블록;을 포함하는, 이미지 처리 장치.

청구항 2

제1 항에 있어서,

상기 초해상화 블록은,

적어도 하나의 제1 잔차(residual) 블록 및 적어도 하나의 제2 잔차 블록을 포함하고,

상기 적어도 하나의 제1 잔차 블록 각각은,

상기 참조 패치 이미지에 대응하는 참조 피처 맵(feature map)에 대해 제1 비트(bit) 수에 기초하여 양자화를 수행하고,

상기 적어도 하나의 제2 잔차 블록 각각은,

상기 쿼리 패치 이미지에 대응하는 쿼리 피처 맵에 대해, 상기 제1 비트 수보다 적은 제2 비트 수에 기초하여 양자화를 수행하는 것을 특징으로 하는, 이미지 처리 장치.

청구항 3

제2 항에 있어서,

상기 적어도 하나의 제1 잔차 블록 각각은,

상기 참조 피처 맵을 연산하기 위해 사용되는 참조 가중치에 대해 상기 제1 비트 수에 기초하여 양자화를 수행하고,

상기 적어도 하나의 제2 잔차 블록 각각은,

상기 쿼리 피처 맵을 연산하기 위해 사용되는 쿼리 가중치에 대해 제3 비트 수에 기초하여 양자화를 수행하는 것을 특징으로 하는, 이미지 처리 장치.

청구항 4

제1 항에 있어서,

상기 클러스터링 블록은,

상기 복수의 패치 이미지들 각각에 대응하는 특징 벡터에 소프트맥스(softmax) 함수를 적용하여 제1 분류 데이

터를 생성하고,

상기 복수의 패치 이미지들 각각에 대응하는 특징 벡터에 전치 소프트맥스(transpose softmax) 함수를 적용하여 제2 분류 데이터를 생성하고,

상기 제1 분류 데이터 및 상기 제2 분류 데이터에 기초하여 상기 각 클러스터의 상기 참조 패치 이미지 및 상기 쿼리 패치 이미지를 구별하는 것을 특징으로 하는, 이미지 처리 장치.

청구항 5

제1 항에 있어서,

상기 이미지 처리 장치는,

대상 클러스터에 포함된 쿼리 패치 이미지들 중 대상 쿼리 패치 이미지에 대해 양자화 에러를 복원하기 위한 적어도 하나의 에러 복원 블록을 포함하고,

상기 적어도 하나의 에러 복원 블록은,

상기 대상 클러스터에 포함된 상기 참조 패치 이미지에 대응하는 보조 피처 맵에 기초하여 상기 대상 쿼리 패치 이미지에 대응하는 대상 쿼리 피처 맵의 양자화 에러를 복원하는 것을 특징으로 하는, 이미지 처리 장치.

청구항 6

제5 항에 있어서,

상기 적어도 하나의 에러 복원 블록은,

상기 보조 피처 맵을 상기 대상 쿼리 피처 맵에 워핑하는 워핑(warping) 연산을 수행하고,

상기 워핑된 보조 피처 맵과 상기 대상 쿼리 피처 맵을 감산하는 감산 연산을 수행하고,

상기 감산 연산의 결과에 양자화 및 활성 함수 연산을 수행하여 복원 쿼리 피처 맵을 생성하고,

상기 복원 쿼리 피처 맵 및 상기 대상 쿼리 피처 맵을 합산하는 합산 연산을 수행하는 것을 특징으로 하는, 이미지 처리 장치.

청구항 7

복수의 패치 이미지들을 획득하는 단계;

상기 복수의 패치 이미지들을 클러스터링하여 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 구별 데이터를 생성하는 단계;

상기 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성하는 단계;

상기 구별 데이터에 기초하여 상기 대상 피처 맵의 피처 맵 유형을 판단하는 단계;

상기 대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트로 상기 대상 피처 맵에 대해 양자화를 수행하는 단계;를 포함하는, 이미지 처리 방법.

청구항 8

제7 항에 있어서,

상기 양자화를 수행하는 단계는,

상기 대상 피처 맵의 피처 맵 유형이 상기 참조 패치 이미지에 대응하는 참조 피처 맵인 경우, 상기 대상 피처 맵에 대해 제1 비트(bit) 수에 기초하여 양자화를 수행하는 단계; 및

상기 대상 피처 맵의 피처 맵 유형이 상기 쿼리 패치 이미지에 대응하는 쿼리 피처 맵인 경우, 상기 대상 피처 맵에 대해 상기 제1 비트 수보다 적은 제2 비트 수에 기초하여 양자화를 수행하는 단계;를 포함하는, 이미지 처리 방법.

청구항 9

제8 항에 있어서,

상기 제2 비트 수에 기초하여 양자화를 수행하는 단계는,

상기 대상 클러스터에 포함된 참조 패치 이미지에 대응하고, 상기 제1 비트 수에 기초하여 양자화가 수행된 보조 피처 맵을 이용하여, 상기 제2 비트 수에 기초하여 양자화가 수행된 상기 대상 피처 맵의 양자화 에러를 복원하는 단계;를 더 포함하는, 이미지 처리 방법.

청구항 10

이미지 처리 장치에 있어서,

하나 이상의 인스트럭션을 저장하는 메모리; 및

하나 이상의 프로세서;를 포함하고, 상기 하나 이상의 프로세서는 상기 하나 이상의 인스트럭션을 실행하여,

복수의 패치 이미지들을 획득하고,

상기 복수의 패치 이미지들을 클러스터링하여 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 구별 데이터를 생성하고,

상기 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성하고,

상기 구별 데이터에 기초하여 상기 대상 피처 맵의 피처 맵 유형을 판단하고,

상기 대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트 수로 상기 대상 피처 맵에 대해 양자화를 수행하는 것을 특징으로 하는, 이미지 처리 장치.

발명의 설명

기 술 분 야

[0001] 본 개시의 기술적 사상은 뉴럴 네트워크 프로세서를 이용한 이미지 처리 동작을 수행하는 이미지 처리 장치 및 이미지 처리 방법에 관한 발명이다.

배 경 기 술

[0002] 최근, 고품질 및 고화질의 사진, 영상 등에 대한 요구가 커짐에 따라 이미지 센서로부터 생성된 이미지 데이터를 뉴럴 네트워크 프로세서(Neural Network Processor)를 이용하여 이미지 처리를 효율적으로 수행할 수 있다. 뉴럴 네트워크에 기초하여 이미지 처리를 위한 딥러닝(deep learning) 또는 기계 학습(machine learning) 등이 구현될 수 있다.

[0003] 뉴럴 네트워크는 인간의 생물학적 신경 세포의 특성을 수학적 표현에 의해 모델링한 것으로, 인간이 가지고 있는 학습이라는 능력을 모방한 알고리즘을 이용한다. 이 알고리즘을 통하여, 뉴럴 네트워크는 입력 데이터와 출력 데이터 사이의 사상(mapping)을 생성할 수 있고, 이러한 사상을 생성하는 능력은 뉴럴 네트워크의 학습 능력이라고 표현될 수 있다. 또한, 뉴럴 네트워크는 학습된 결과에 기초하여, 학습에 이용되지 않았던 입력 데이터에 대하여, 올바른 출력 데이터를 생성할 수 있는 일반화 능력을 가진다.

[0004] 심층 신경망(예를 들어, 층이 깊은 컨볼루션 뉴럴 네트워크(CNN: Convolution Neural Network))을 이용하여, 저해상도 이미지로부터 고해상도 이미지를 생성하는 초해상화(super-resolution)와 같은 이미지 처리가 수행될 수 있다. 이미지 데이터를 이미지 처리 시, 연산량을 최소화하면서도 고품질의 이미지를 출력하는 기술이 요구되고 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명의 기술적 사상이 해결하려는 과제는 뉴럴 네트워크 프로세서를 이용하여 참조 패치 이미지 및 쿼리 패치 이미지를 구별하고, 참조 패치 이미지 및 쿼리 패치 이미지에 대해 서로 상이한 비트 수로 양자화를 수행함

으로써, 뉴럴 네트워크 프로세서의 연산량을 감소시키기 위한 이미지 처리 장치 및 이미지 처리 방법을 제공한다.

과제의 해결 수단

[0006] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 개시의 제1 측면은, 입력 이미지를 복수의 패치(patch) 이미지들로 분할하는 프리-프로세서 및 상기 복수의 패치 이미지들 각각에 대해 이미지 처리 동작을 수행하여 출력 이미지를 생성하는 뉴럴 네트워크 프로세서를 포함하고, 상기 뉴럴 네트워크 프로세서는, 상기 복수의 패치 이미지들 중 클러스터의 수에 따라 상기 복수의 패치 이미지들을 클러스터링하고, 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조(reference) 패치 이미지 및 쿼리(query) 패치 이미지로 구별하는 클러스터링 블록 및 상기 각 클러스터의 상기 참조 패치 이미지인지 및 상기 쿼리 패치 이미지인지 여부에 따라, 상기 참조 패치 이미지 및 상기 쿼리 패치 이미지에 대해 서로 상이한 비트(bit) 수로 양자화를 수행하는 초해상화 블록을 포함하는 이미지 처리 장치를 제공할 수 있다.

[0007] 또한, 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 개시의 제2 측면은, 복수의 패치 이미지들을 획득하는 단계, 상기 복수의 패치 이미지들을 클러스터링하여 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 구별 데이터를 생성하는 단계, 상기 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성하는 단계, 상기 구별 데이터에 기초하여 상기 대상 피처 맵의 피처 맵 유형을 판단하는 단계, 상기 대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트로 상기 대상 피처 맵에 대해 양자화를 수행하는 단계를 포함하는, 이미지 처리 방법을 제공할 수 있다.

[0008] 또한, 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 개시의 제3 측면은, 하나 이상의 인스트럭션을 저장하는 메모리 및 하나 이상의 프로세서를 포함하고, 상기 하나 이상의 프로세서는 상기 하나 이상의 인스트럭션을 실행하여, 복수의 패치 이미지들을 획득하고, 상기 복수의 패치 이미지들을 클러스터링하여 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 구별 데이터를 생성하고, 상기 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성하고, 상기 구별 데이터에 기초하여 상기 대상 피처 맵의 피처 맵 유형을 판단하고, 상기 대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트 수로 상기 대상 피처 맵에 대해 양자화를 수행하는 것을 특징으로 하는 이미지 처리 장치를 제공할 수 있다.

발명의 효과

[0009] 본 개시의 기술적 사상에 따른 이미지 처리 장치는, 대상 패치 이미지가 참조 패치 이미지인지 쿼리 패치 이미지인지 여부에 따라, 상이한 비트 수로 대상 패치 이미지를 양자화함으로써, 뉴럴 네트워크의 연산량이 감소될 수 있고, 뉴럴 네트워크 프로세서의 성능이 향상될 수 있다.

[0010] 또한, 이미지 처리 장치는 참조 피처 맵을 이용하여 쿼리 피처 맵의 양자화 에러를 복원함으로써, 양자화 에러가 감소되고 뉴럴 네트워크 프로세서의 성능이 향상될 수 있다. 이에 따라, 이미지 처리 동작 시 고해상도 및 고품질의 이미지가 생성될 수 있다.

[0011] 본 개시의 예시적 실시예들에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 아니하며, 언급되지 아니한 다른 효과들은 이하의 설명으로부터 본 개시의 예시적 실시예들이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 도출되고 이해될 수 있다. 즉, 본 개시의 예시적 실시예들을 실시함에 따른 의도하지 아니한 효과들 역시 본 개시의 예시적 실시예들로부터 당해 기술분야의 통상의 지식을 가진 자에 의해 도출될 수 있다.

도면의 간단한 설명

[0012] 도 1은 본 개시의 예시적인 실시 예에 따른 전자 장치를 나타내는 블록도이다.

도 2는 본 발명의 예시적인 실시예에 따른 뉴럴 네트워크 구조를 설명하기 위한 도면이다.

도 3은 본 개시의 예시적인 실시예에 따른 이미지 처리 장치를 나타내는 블록도이다.

도 4는 일 실시예에 따른 뉴럴 네트워크를 설명하기 위한 도면이다.

도 5는 본 개시의 예시적인 실시예에 따른 클러스터링 블록을 설명하기 위한 도면이다.

도 6은 본 개시의 예시적인 실시예에 따른 구별 블록의 동작을 설명하기 위한 도면이다.

도 7은 본 개시의 예시적인 실시예에 따른 초해상화 블록을 나타내는 도면이다.

도 8은 본 개시의 예시적인 실시예에 따른 잔차 블록을 설명하기 위한 도면이다.

도 9a는 본 개시의 예시적인 실시예에 따른 에러 복원 블록을 설명하기 위한 도면이다.

도 9b는 본 개시의 예시적인 실시예에 따른 에러 복원 블록이 복수 개인 경우를 설명하기 위한 도면이다.

도 10은 본 개시의 예시적 실시예에 따른 에러 복원 블록의 구조를 설명하기 위한 도면이다.

도 11은 본 개시의 예시적 실시 예에 따른 이미지 처리 방법을 나타내는 흐름도이다.

도 12는 본 개시의 예시적인 실시예에 따른 이미지 처리 장치가 양자화를 수행하는 방법을 나타내는 흐름도이다.

도 13은 본발명의 일 실시예에 따른 이미지 처리 장치를 설명하기 위한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하, 첨부한 도면을 참조하여 본 발명의 실시 예에 대해 상세히 설명한다.

[0014] 도 1은 본 개시의 예시적인 실시 예에 따른 전자 장치를 나타내는 블록도이다.

[0015] 본 개시의 예시적 실시예에 따른 전자 장치(10)는 뉴럴 네트워크를 기초로 입력 이미지에 대해 이미지 처리 동작을 수행하고, 출력 이미지를 생성할 수 있다. 전자 장치(10)는 스마트폰(smartphone), 태블릿 PC(tablet personal computer), 이동 전화기(mobile phone), 화상 전화기, 전자북 리더기(e-book reader), 데스크탑 PC(desktop personal computer), 랩탑 PC(laptop personal computer), 넷북 컴퓨터(netbook computer), 워크스테이션(workstation), 서버, PDA(personal digital assistant), PMP(portable multimedia player), MP3 플레이어, 모바일 의료기기, 카메라(camera), 또는 웨어러블 장치(wearable device) 중 적어도 하나를 포함할 수 있다. 이 뿐 아니라, 전자 장치(10)는 스마트 가전 제품(smart home appliance)을 포함할 수 있다. 스마트 가전 제품은, 예를 들어, 텔레비전, DVD(digital video disk) 플레이어, 오디오, 냉장고, 에어컨, 청소기, 오븐, 전자레인지, 세탁기, 공기 청정기, 셋톱 박스(set-top box), 홈 오토메이션 컨트롤 패널(home automation control panel), 보안 컨트롤 패널(security control panel), TV 박스, 게임 콘솔, 전자 사전, 전자 키, 캠코더, 또는 전자 액자 중 적어도 하나를 포함할 수 있다.

[0016] 또한, 전자 장치(10)는 각종 의료기기를 포함할 수 있다. 의료기기는, 예를 들어, 각종 휴대용 의료측정기기(혈당 측정기, 심박 측정기, 혈압 측정기 또는 체온 측정기 등), MRA(magnetic resonance angiography), MRI(magnetic resonance imaging), CT(computed tomography), 촬영기 또는 초음파기 등을 포함할 수 있다. 예를 들어, 전자 장치(10)는 어플리케이션 프로세서일 수 있다. 어플리케이션 프로세서는 다양한 종류의 연산 처리를 수행할 수 있다. 도 1에는 도시되어 있지 않으나, 전자 장치(10)는 뉴럴 네트워크를 이용 하여 처리할 연산을 분담하는 NPU(Neural processing unit)를 더 포함할 수 있다.

[0017] 도 1을 참조하면, 전자 장치(10)는 이미지 처리 장치(100), CPU(Central Proccessing Unit)(200), RAM(random access memory)(300), 카메라 모듈(400), 메모리(500), 디스플레이(600), 및 시스템 버스(700)를 포함할 수 있다. 실시예에 따라, 전자 장치(10)는 도 1에 도시된 구성 요소들 외에 다른 범용적인 구성 요소들을 더 포함할 수 있다. 예를 들어, 전자 장치(10)는 입출력 모듈, 보안 모듈, 전력 제어 장치 등을 더 포함할 수 있으며, 또한 다양한 종류의 프로세서들을 더 포함할 수 있다. 또한, 실시예에 따라, 도 1의 구성 요소들 중 적어도 하나가 전자 장치(10)에서 생략될 수도 있다. 전자 장치(10)의 구성들은 버스(700)를 통해 서로 통신할 수 있다.

[0018] 실시예들에 있어서, 전자 장치(10)의 구성들 중 일부 또는 전부는 하나의 반도체 칩에 형성될 수 있다. 예컨대, 전자 장치(10)는 시스템-온 칩(System On Chip; SoC)으로서 구현될 수 있으며, 일부 실시예에 있어서, 이미지 칩 등으로 지칭될 수 있다.

[0019] 이미지 처리 장치(100)는 입력 이미지에 대해 이미지 처리를 수행하여 출력 이미지를 생성할 수 있다. 이미지 처리 장치(100)는 뉴럴 네트워크 프로세서(110)를 이용하여 입력 이미지에 대해 이미지 처리를 수행하고 출력 이미지를 생성할 수 있다. 입력 이미지는 입력 이미지 데이터, 입력 데이터 등으로 지칭될 수도 있다. 뉴럴 네트워크 프로세서(110)는 뉴럴 네트워크를 훈련(train)(또는 학습(learn))하거나, 뉴럴 네트워크를 이용하여 입력 데이터를 분석함으로써 입력 데이터에 포함된 정보를 추론(inference)할 수 있다. 뉴럴 네트워크 프로세서

(110)는 추론된 정보를 기초로 상황 판단을 하거나 또는 뉴럴 네트워크 프로세서(110)가 탑재되는 전자 장치(10)의 구성들을 제어할 수 있다.

[0020] 뉴럴 네트워크 프로세서(110)는 수신되는 입력 이미지를 기초로 뉴럴 네트워크 연산을 수행할 수 있다. 나아가, 뉴럴 네트워크 프로세서(110)는 뉴럴 네트워크 연산의 수행 결과를 기초로 정보 신호를 생성할 수 있다. 뉴럴 네트워크 프로세서(110)는 뉴럴 네트워크 연산 가속기, 코프로세서(coprocessor), DSP(Digital Signal Processor), ASIC(Application Specific Integrated Circuit), FPGA(Field-programmable gate array), GPU(Graphics processing unit), NPU(Neural processing unit), TPU(Tensor Processing Unit) 및 MPSoC(Multi-Processor System-on-Chip) 등으로 구현될 수 있다.

[0021] 뉴럴 네트워크 프로세서(110)는 ANN(Artificial Neural Network), CNN(Convolution Neural Network), R-CNN(Region with Convolution Neural Network), RPN(Region Proposal Network), RNN(Recurrent Neural Network), S-DNN(Stacking-based deep Neural Network), S-SDNN(State-Space Dynamic Neural Network), Deconvolution Network, DBN(Deep Belief Network), RBM(Restricted Boltzman Machine), Fully Convolutional Network, LSTM(Long Short-Term Memory) Network, Classification Network, Plain Residual Network, Dense Network, Hierarchical Pyramid Network, Fully Convolutional Network 중 적어도 하나에 기초한 뉴럴 네트워크 프로세서(110)일 수 있다. 한편, 뉴럴 네트워크 프로세서(110)의 종류는 전술한 예에 한하지 않는다.

[0022] 이미지 처리 장치(100)는 입력 이미지를 수신할 수 있다. 예시적으로, 이미지 처리 장치(100)는 카메라 모듈(400)로부터 입력 이미지를 수신할 수 있다. 구체적으로, 이미지 처리 장치(100)는 카메라 모듈(400)의 이미지 센서로부터 생성된 입력 이미지를 수신하고, 입력 이미지에 대한 이미지 처리 동작들을 수행하여 출력 이미지를 생성할 수 있다. 다만, 이에 반드시 제한되는 것은 아니다. 이미지 처리 장치(100)는 전자 장치(10)에 기 저장된 입력 이미지에 대해 이미지 처리 동작을 수행할 수도 있고, 전자 장치(10)의 외부로부터 수신되는 입력 이미지에 대해 이미지 처리 동작을 수행할 수도 있다.

[0023] 본 개시의 예시적 실시예에 따른 이미지 처리 장치(100)는 카메라 모듈(400) 또는 메모리(500)로부터 이미지 데이터를 수신하고, 이에 기반하여 뉴럴 네트워크 연산을 수행할 수 있다. 이미지 처리 장치(100)는 뉴럴 네트워크 연산을 통하여 정의된 이미지 처리 동작을 수행할 수 있다.

[0024] 예시적 실시예에 따른 뉴럴 네트워크 프로세서(110)는 입력 이미지에 대해 이미지 처리 동작을 수행하도록 학습될 수 있다. 뉴럴 네트워크 프로세서(110)는 입력 이미지에 대해 이미지 처리 동작을 수행하여 출력 이미지를 생성할 수 있다. 일 실시예에서, 이미지 처리 동작은 입력 이미지의 고해상도 이미지를 생성하는 초해상화(super-resolution) 동작일 수 있다. 입력 이미지는 노이즈를 포함하는 이미지이거나, 저해상도 이미지일 수 있다. 출력 이미지는 입력 이미지보다 고해상도 이미지일 수 있고, 입력 이미지보다 화질이 개선된 이미지일 수 있다.

[0025] 실시예에 따라, 이미지 처리 장치(100)는 배드 픽셀 보정(Bad Pixel Correction, BPC) 동작, 렌즈 쉐이딩 보정(Lens Shading Correction, LSC) 동작, 크로스 토크 보정(X-talk correction) 동작, 리모자이크(Remosaic) 동작, 디모자이크(Demosaic) 동작, 디노이즈(Denoise) 동작 등 다양한 이미지 처리 동작들을 더 수행할 수도 있다. 한편, 이미지 처리 동작의 종류는 전술한 예에 한하지 않는다.

[0026] 이미지 처리 장치(100)는 입력 이미지를 수신하고, 이를 복수의 이미지들로 분할할 수 있다. 이하에서는, 이미지 처리 장치(100)가 분할한 입력 이미지를 패치 이미지로 지칭할 수 있다. 이미지 처리 장치(100)는 입력 이미지를 특정 크기의 패치 이미지들로 분할할 수 있다. 예시적으로, 패치 이미지들은 서로 동일한 크기일 수도 있고, 상이한 크기일 수도 있다. 복수의 패치 이미지들은 뉴럴 네트워크 프로세서(110)로 입력될 수 있다. 뉴럴 네트워크 프로세서(110)는 패치 이미지들 각각에 대해 이미지 처리 동작을 수행할 수 있다.

[0027] 뉴럴 네트워크 프로세서(110)는 패치 이미지에 대해 뉴럴 네트워크 연산을 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 패치 이미지들 각각으로부터 피처(feature)를 추출할 수 있고, 이에 기반하여 뉴럴 네트워크 연산을 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 고품질 및 고해상도의 이미지를 생성하기 위한 이미지 처리 동작을 수행하도록 훈련된 뉴럴 네트워크를 이용할 수 있다.

[0028] 뉴럴 네트워크 프로세서(110)는 초해상화 동작을 수행하기 위해 클러스터링 동작을 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 복수의 패치 이미지들 중 클러스터의 수에 따라 복수의 패치 이미지들을 클러스터링하여 복수의 패치 이미지들 각각이 참조 패치 이미지인지 쿼리 패치 이미지인지 구별할 수 있다. 뉴럴 네트워크 프로세서(110)는 클러스터의 수에 따라 복수의 패치 이미지들을 클러스터링할 수 있다. 클러스터의 수는 기 설정되

거나 학습에 따라 정해질 수 있다. 뉴럴 네트워크 프로세서(110)는 클러스터 수에 따라 유사한 패치 이미지들끼리 클러스터를 형성하도록 클러스터링할 수 있다.

[0029] 뉴럴 네트워크 프로세서(110)는 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조 패치 이미지 및 쿼리 패치 이미지로 구별할 수 있다. 참조 패치 이미지는 각 클러스터를 대표하는 이미지를 의미하고, 쿼리 패치 이미지는 각 클러스터에서 참조 패치 이미지가 아닌 패치 이미지를 의미할 수 있다. 뉴럴 네트워크 프로세서(110)는 각 클러스터의 특징을 가장 잘 나타내는 패치 이미지를 각 클러스터의 참조 패치 이미지로 구별할 수 있다. 예시적으로, 제1 클러스터에 제1 패치 이미지, 제2 패치 이미지, 제3 패치 이미지가 포함된 경우, 뉴럴 네트워크 프로세서(110)는 제1 클러스터의 특징을 가장 잘 나타내는 제1 패치 이미지를 참조 패치 이미지로, 제2 패치 이미지 및 제3 패치 이미지를 쿼리 패치 이미지로 구별할 수 있다.

[0030] 패치 이미지에 대해 이미지 처리 동작을 수행하기 위해 파라미터들이 이용될 수 있다. 뉴럴 네트워크 프로세서(110)는 뉴럴 네트워크 연산을 수행하기 위해 파라미터들을 이용할 수 있다. 파라미터들은 패치 이미지에 대해 이미지 처리 동작을 수행하여 출력 이미지를 생성하는 학습을 통해 획득될 수 있다. 예시적으로, 파라미터들은 패치 이미지에 대해 초해상화 동작을 수행하여 출력 이미지를 생성하는 학습을 통해 획득될 수 있다. 파라미터는 뉴럴 네트워크 연산을 수행하기 위해 이용되는 것으로, 파라미터들은 예를 들어 뉴럴 네트워크의 입/출력 액티베이션들, 웨이트들, 바이어스들 등 뉴럴 네트워크에 입/출력되는 다양한 종류의 데이터를 포함할 수 있다. 예시적으로, 파라미터는 이미지 처리 장치(100)의 내부 메모리에 저장될 수도 있고, 메모리(500)에 저장될 수도 있다. 메모리(500)에 저장된 파라미터는 이미지 처리 장치(100)에 제공될 수 있다.

[0031] 뉴럴 네트워크 프로세서(110)는 패치 이미지들 각각에 대해 양자화를 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 패치 이미지들 각각의 이미지 처리를 수행하기 위해, 획득된 파라미터들의 양자화를 수행할 수 있다. 파라미터 양자화는, 부동 소수점(floating point) 타입의 파라미터를, 고정(fixed) 소수점 타입의 파라미터로 변환하는 것을 의미할 수 있다. 예시적으로, 뉴럴 네트워크 프로세서(110)는 패치 이미지들 각각에 대한 뉴럴 네트워크 연산을 수행하기 위한 액티베이션들 및 웨이트들의 양자화를 수행할 수 있다. 이미지에 대해 양자화를 수행하는 것은, 특정 이미지를 고해상도의 이미지로 이미지 처리 하기 위해 사용되는 파라미터를 양자화하는 것을 의미할 수 있다.

[0032] 일 실시예에서, 뉴럴 네트워크 프로세서(110)는 각 클러스터의 참조 패치 이미지인지 및 쿼리 패치 이미지인지 여부에 따라, 참조 패치 이미지 및 쿼리 패치 이미지에 대해 서로 상이한 비트 수로 양자화를 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 패치(patch) 이미지에 대응하는 피처 맵(feature map)에 대해 양자화를 수행할 수 있다. 예시적으로, 뉴럴 네트워크 프로세서(110)는 이미지 처리의 대상인 대상 패치 이미지가 참조 패치 이미지인 경우, 대상 패치 이미지에 대응하는 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 예를 들어, 뉴럴 네트워크 프로세서(110)는 제1 클러스터의 제1 패치 이미지가 참조 패치 이미지에 해당하는 경우, 제1 패치 이미지에 대응하는 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다.

[0033] 예시적으로, 뉴럴 네트워크 프로세서(110)는 이미지 처리의 대상인 대상 패치 이미지가 쿼리 패치 이미지인 경우, 대상 패치 이미지에 대응하는 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 예를 들어, 뉴럴 네트워크 프로세서(110)는 제1 클러스터의 제2 패치 이미지가 쿼리 패치 이미지에 해당하는 경우, 제2 패치 이미지에 대응하는 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다.

[0034] 저해상도의 이미지를 고해상도의 이미지로 이미지 처리하기 위해 뉴럴 네트워크 연산이 수행되는 경우, 많은 연산량이 필요하다. 이미지 처리 장치는 뉴럴 네트워크 연산 수행 시, 참조 패치 이미지에 대해 많은 비트 수로 양자화를 수행하고, 쿼리 패치 이미지에 대해 상대적으로 적은 비트 수로 양자화를 수행함으로써, 연산량을 감소시킬 수 있고, 이미지 처리 속도가 향상될 수 있다.

[0035] 카메라 모듈(400)은 전자 장치(10) 외부의 피사체(또는 객체)를 촬영하고, 이미지 데이터를 생성할 수 있다. 예를 들어, 카메라 모듈(400)은 이미지 센서를 포함할 수 있다. 이미지 센서는 광학 렌즈(미도시)에 의해 피사체의 광학적 신호를 전기적 신호로 변환할 수 있다. 이를 위해, 이미지 센서는 복수의 픽셀들이 2차원적으로 배열된 픽셀 어레이를 포함할 수 있다. 예를 들어, 복수의 픽셀들 각각에는 복수의 기준색들 중 하나의 색이 할당될 수 있다. 예를 들어, 복수의 기준색들은 RGB(red, green, blue), 또는 RGBW(red, green, blue, white)를 포함할 수 있다.

[0036] 카메라 모듈(400)은 이미지 센서를 이용하여 입력 이미지를 생성할 수 있다. 입력 이미지는 이미지 데이터, 이미지 프레임 및 프레임 데이터와 같이 다양하게 지칭될 수 있다. 입력 이미지는 이미지 처리 장치(100)에 입력

데이터로서 제공되거나, 메모리(500)에 저장될 수 있다. 메모리(500)에 저장된 입력 이미지는 이미지 처리 장치(100)에 입력 데이터로서 제공될 수 있다.

[0037] CPU(200)는 전자 장치(10)의 전반적인 동작을 제어한다. CPU(200)는 하나의 프로세서 코어(Single Core)를 포함하거나, 복수의 프로세서 코어들(Multi-Core)을 포함할 수 있다. CPU(200)는 메모리(500)와 같은 저장 영역에 저장된 프로그램들 및/또는 데이터를 램(300)을 이용하여 처리 또는 실행할 수 있다.

[0038] 메모리(500)는 휘발성 메모리(volatile memory) 또는 비휘발성 메모리(nonvolatile memory) 중 적어도 하나를 포함할 수 있다. 비휘발성 메모리는 ROM(Read Only Memory), PROM(Programmable ROM), EPROM(Electrically Programmable ROM), EEPROM(Electrically Erasable and Programmable ROM), 플래시 메모리, PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM) 등을 포함한다. 휘발성 메모리는 DRAM(Dynamic RAM), SRAM(Static RAM), SDRAM(Synchronous DRAM), PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM), FeRAM(Ferroelectric RAM) 등을 포함한다. 실시예에 있어서, 메모리(500)는 HDD(Hard Disk Drive), SSD(Solid State Drive), CF(compact flash) 카드, SD(secure digital) 카드, Micro-SD(micro secure digital) 카드, Mini-SD(mini secure digital) 카드, xD(extreme digital) 카드 또는 Memory Stick 중 적어도 하나를 포함할 수 있다.

[0039] 디스플레이(600)는 이미지 처리 장치(100)로부터 수신한 이미지 데이터를 기초로 사용자에게 각종 컨텐츠(예를 들어, 텍스트, 이미지, 비디오, 아이콘, 또는 심볼 등)를 표시할 수 있다. 예를 들어, 디스플레이(600)는 액정 디스플레이(LCD), 발광 다이오드(LED) 디스플레이, 유기 발광 다이오드(OLED) 디스플레이, 또는 마이크로 전자기계 시스템(MEMS) 디스플레이, 또는 전자종이(electronic paper) 디스플레이를 포함할 수 있다. 디스플레이(600)는 이미지를 표시하기 위해 복수의 픽셀들이 매트릭스 형태로 배치된 픽셀 어레이를 포함할 수 있다.

[0040] 도 2는 본 발명의 예시적인 실시예에 따른 뉴럴 네트워크 구조를 설명하기 위한 도면이다. 이는, 도 2는 뉴럴 네트워크(NN)의 구조를 개략적으로 도시한 것으로, 도 1의 뉴럴 네트워크 프로세서(110)에 적용될 수 있다. 이미지 처리 장치는 뉴럴 네트워크(NN)를 이용하여 패치 이미지들을 클러스터링할 수 있다. 이미지 처리 장치는 뉴럴 네트워크(NN)를 이용하여 패치 이미지에 대해 양자화할 수 있다.

[0041] 도 2를 참조하면, 뉴럴 네트워크(NN)는 복수의 레이어들(L1 내지 Ln)을 포함할 수 있다. 이와 같은 멀티-레이어드 구조의 뉴럴 네트워크는 딥 뉴럴 네트워크(deep neural network; DNN) 또는 딥 러닝 구조(deep learning architecture)로 지칭될 수 있다. 복수의 레이어들(L1 내지 Ln) 각각은 선형 레이어 또는 비선형 레이어일 수 있으며, 실시예에 있어서, 적어도 하나의 레이어 및 적어도 하나의 비선형 레이어가 결합되어 하나의 레이어로 지칭될 수도 있다. 예시적으로, 선형 레이어는 컨볼루션 레이어(convolution), 풀리 커넥티드(fully connected) 레이어를 포함할 수 있으며, 비선형 레이어는 풀링(pooling) 레이어, 액티베이션 레이어를 포함할 수 있다.

[0042] 예시적으로, 제1 레이어(L1)는 컨볼루션 레이어이고, 제2 레이어(L2)는 풀링(pooling) 레이어이고, 제n 레이어(Ln)는 출력 레이어로서 풀리 커넥티드(fully connected) 레이어일 수 있다. 뉴럴 네트워크(NN)는 활성(activation) 레이어를 더 포함할 수 있으며, 다른 종류의 연산을 수행하는 레이어를 더 포함할 수 있다.

[0043] 복수의 레이어들(L1 내지 Ln) 각각은 입력되는 입력 이미지 또는 이전 레이어에서 생성된 피처 맵을 입력 피처 맵으로서 수신하고, 입력 피처 맵을 연산하여 출력 피처 맵을 생성할 수 있다. 이때, 피처 맵은 입력 데이터의 다양한 특징이 표현된 데이터를 의미한다. 피처 맵들(FM1, FM2, FM3, FMn)은 예컨대 복수의 피처 값들을 포함하는 2차원 매트릭스 또는 3차원 매트릭스(또는 텐서(tensor)라고 함) 형태를 가질 수 있다. 피처 맵들(FM1, FM2, FM3, FMn)은 너비(W)(또는 칼럼이라고 함), 높이(H)(또는 로우라고 함) 및 깊이(D)를 가지며, 이는 좌표상의 x축, y축 및 z축에 각각 대응할 수 있다. 이때, 깊이(D)는 채널 수로 지칭될 수 있다.

[0044] 제1 레이어(L1)는 제1 피처 맵(FM1)을 웨이트 맵(WM)과 컨볼루션하여 제2 피처 맵(FM2)을 생성할 수 있다. 웨이트 맵(WM)은 복수의 웨이트 값들을 포함하는 2차원 매트릭스 또는 3차원 매트릭스 형태를 가질 수 있다. 웨이트 맵(WM)은 커널로 지칭될 수 있다. 웨이트 맵(WM)은 제1 피처 맵(FM1)을 필터링할 수 있으며, 필터 또는 커널로 지칭될 수 있다. 웨이트 맵(WM)의 깊이, 즉 채널 개수는 제1 피처 맵(FM1)의 깊이, 즉 채널 개수와 동일하며, 웨이트 맵(WM)과 제1 피처 맵(FM1)의 동일한 채널끼리 컨볼루션될 수 있다. 웨이트 맵(WM)이 제1 피처 맵(FM1)을 슬라이딩 윈도로하여 횡단하는 방식으로 시프트된다. 각 시프트동안, 웨이트 맵(WM)에 포함되는 웨이트들 각각이 제1 피처 맵(FM1)과 중첩된 영역에서의 모든 피처값과 곱해지고 더해질 수 있다. 제1 피처 맵(FM1)과 웨이트 맵(WM)이 컨볼루션 됨에 따라, 제2 피처 맵(FM2)의 하나의 채널이 생성될 수 있다. 도 2에는 하나의 웨이트 맵(WM)이 표시되었으나, 실질적으로는 복수개의 웨이트 맵이 제1 피처 맵(FM1)과 컨볼루션됨으로써, 제2 피처

맵(FM2)의 복수개의 채널이 생성될 수 있다. 다시 말해, 제2 피처 맵(FM2)의 채널의 수는 웨이트 맵의 개수에 대응할 수 있다.

[0045] 제2 레이어(L2)는 풀링을 통해 제2 피처 맵(FM2)의 공간적 크기(spatial size)를 변경함으로써, 제3 피처 맵(FM3)을 생성할 수 있다. 풀링은 샘플링 또는 다운-샘플링으로 지칭될 수 있다. 2 차원의 풀링 윈도우(WD)가 풀링 윈도우(WD)의 사이즈 단위로 제2 피처 맵(FM2) 상에서 쉬프트되고, 풀링 윈도우(WD)와 중첩되는 영역의 피처 값들 중 최대값(또는 피처값들의 평균값)이 선택될 수 있다. 이에 따라, 제2 피처 맵(FM2)으로부터 공간적 사이즈가 변경된 제3 피처 맵(FM3)이 생성될 수 있다. 제3 피처 맵(FM3)의 채널과 제2 피처 맵(FM2)의 채널 개수는 동일하다.

[0046] 실시예에 따라, 제2 레이어(L2)는 샘플링 레이어 또는 풀링 레이어에 한정되지 않는다. 예를 들어, 제2 레이어(L2)는 제1 레이어(L1)와 유사한 컨볼루션 레이어가 될 수도 있다. 제2 레이어(L2)는 제2 피처 맵(FM2)을 웨이트 맵과 컨볼루션하여 제3 피처 맵(FM3)을 생성할 수 있다. 이 경우, 제2 레이어(L2)에서 컨볼루션 연산을 수행한 웨이트 맵은 제1 레이어(L1)에서 컨볼루션 연산을 수행한 웨이트 맵(WM)과 다를 수 있다. 제1 레이어(L1) 및 제2 레이어(L2)를 포함한 복수의 레이어들을 거쳐 제N 레이어에서 제N 피처 맵을 생성할 수 있다.

[0047] 도 3은 본 개시의 예시적인 실시예에 따른 이미지 처리 장치를 나타내는 블록도이다. 도 3의 이미지 처리 장치(100) 및 뉴럴 네트워크 프로세서(110)는 도 1의 이미지 처리 장치(100) 및 뉴럴 네트워크 프로세서(110)에 대응되므로 중복되는 내용은 생략한다.

[0048] 이미지 처리 장치(100)는 이미지를 촬영하고, 촬영된 이미지를 이미지 처리하거나 또는 저장되거나 수신한 이미지를 이미지 처리하는 동작을 수행할 수 있다. 또한, 이미지 처리 장치(100)는 전자 기기 또는 차량, 가구, 제조 설비, 도어, 각종 계측 기기 등에 부품으로서 구비되는 전자 기기에 탑재될 수 있다.

[0049] 도 3을 참조하면, 이미지 처리 장치(100)는 프리-프로세서(120) 및 뉴럴 네트워크 프로세서(110)를 포함할 수 있다. 프리-프로세서(120)는 입력 이미지(IDT)를 수신하고, 입력 이미지(IDT)에 대해 패치 이미지(PI)들을 생성할 수 있다. 프리-프로세서(120)는 입력 이미지(IDT)를 분할하여 패치 이미지(PI)들을 생성할 수 있다.

[0050] 프리-프로세서(120)는 입력 이미지(IDT)를 특정 크기의 패치 이미지(PI)들로 분할할 수 있다. 예시적으로, 패치 이미지들은 서로 동일한 크기일 수도 있고, 상이한 크기일 수도 있다. 예를 들어, 프리-프로세서(120)는 입력 이미지(IDT)를 6개의 패치 이미지(PI)들로 분할할 수 있다. 복수의 패치 이미지(PI)들은 뉴럴 네트워크 프로세서(110)로 입력될 수 있다.

[0051] 뉴럴 네트워크 프로세서(110)는 패치 이미지(PI)들 각각에 대해 이미지 처리 동작을 수행할 수 있다. 뉴럴 네트워크 프로세서(110)는 패치 이미지(PI)에 대해 뉴럴 네트워크 연산을 수행할 수 있다. 도 3을 참조하면, 뉴럴 네트워크 프로세서(110)에 적용되는 뉴럴 네트워크는 클러스터링 블록(111) 및 초해상화 블록(112)을 포함할 수 있다. 실시예에 따라, 프리-프로세서(120) 및 뉴럴 네트워크 프로세서(110)는 단일 또는 복수의 반도체 칩으로 구현될 수 있다.

[0052] 뉴럴 네트워크 프로세서(110)는 고해상도의 이미지를 생성하기 위해 클러스터링 동작을 수행할 수 있다. 상세하게는, 클러스터링 블록(111)은 패치 이미지(PI)들을 수신할 수 있다. 클러스터링 블록(111)은 복수의 패치 이미지(PI)들을 클러스터링하여 복수의 패치 이미지(PI)들 각각이 참조 패치 이미지인지 쿼리 패치 이미지인지 구별할 수 있다. 클러스터링 블록(111)은 클러스터 수에 따라 유사한 패치 이미지들끼리 클러스터를 형성하도록 클러스터링할 수 있다. 예시적으로, 클러스터링 블록(111)은 제1 패치 이미지 및 제2 패치 이미지를 제1 클러스터로 클러스터링하고, 제3 패치 이미지 및 제4 패치 이미지를 제2 클러스터로 클러스터링하고, 제5 패치 이미지 및 제6 패치 이미지를 제3 클러스터로 클러스터링할 수 있다. 다만, 이는 일 예시일 뿐 이에 반드시 제한되는 것은 아니다.

[0053] 클러스터링 블록(111)은 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조 패치 이미지 및 쿼리 패치 이미지로 구별할 수 있다. 예시적으로, 클러스터링 블록(111)은 참조 패치 이미지인지 쿼리 패치 이미지인지 구별하는 구별 데이터를 생성할 수 있다. 클러스터링 블록(111)은 각 클러스터의 특징을 가장 잘 나타내는 패치 이미지(PI)를 각 클러스터의 참조 패치 이미지로 구별할 수 있다. 예시적으로, 클러스터링 블록(111)은 제1 클러스터의 특징을 가장 잘 나타내는 제1 패치 이미지를 제1 클러스터의 참조 패치 이미지로 구별할 수 있다. 클러스터링 블록(111)은 각 클러스터에서 참조 패치 이미지가 아닌 패치 이미지를 쿼리 패치 이미지로 구별할 수 있다. 클러스터링 블록(111)은 제2 패치 이미지를 제1 클러스터의 쿼리 패치 이미지로 구별할 수 있다.

[0054] 뉴럴 네트워크 프로세서(110)는 고해상도의 이미지를 생성하기 위해 양자화 동작을 포함하는 이미지 처리 동작

을 수행할 수 있다. 초해상화 블록(112)은 패치 이미지(PI)들 각각에 대해 이미지 처리를 위한 다양한 연산을 수행할 수 있다. 초해상화 블록(112)은 패치 이미지(PI)들 각각에 대해 양자화를 수행할 수 있다.

[0055] 상세하게는, 초해상화 블록(112)은 패치 이미지(PI)에 대응하는 피처 맵을 수신할 수 있다. 초해상화 블록(112)은 피처 맵들에 대해 이미지 처리를 위한 다양한 연산을 수행할 수 있다. 초해상화 블록(112)은 패치 이미지(PI)들 각각에 대응하는 피처 맵에 대한 이미지 처리 동작을 수행하기 위해, 파라미터들의 양자화를 수행할 수 있다. 예시적으로, 초해상화 블록(112)은 패치 이미지들 각각에 대한 뉴럴 네트워크 연산을 수행하기 위한 액티베이션들 및 웨이트들의 양자화를 수행할 수 있다.

[0056] 초해상화 블록(112)은 각 클러스터의 참조 패치 이미지인지 및 쿼리 패치 이미지인지 여부에 따라, 참조 패치 이미지 및 쿼리 패치 이미지에 대해 서로 상이한 비트 수로 양자화를 수행할 수 있다. 일 실시예에서, 초해상화 블록(112)은 참조 패치 이미지에 대응하는 피처 맵에 대해 상대적으로 많은 비트 수로 양자화를 수행할 수 있다. 예시적으로, 초해상화 블록(112)은 이미지 처리의 대상인 대상 패치 이미지가 참조 패치 이미지인 경우, 대상 패치 이미지에 대응하는 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 이는, 대상 패치 이미지에 대응하는 피처 맵에 대해 제1 비트의 정밀도로 양자화가 수행되는 것을 의미할 수 있다.

[0057] 초해상화 블록(112)은 참조 패치 이미지에 대응하는 참조 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 이미지 처리의 대상인 대상 패치 이미지가 포함된 클러스터는 대상 클러스터로 지칭될 수 있다. 대상 패치 이미지가 참조 패치 이미지인 경우 대상 참조 패치 이미지로 지칭되고, 대상 참조 패치 이미지에 대응하는 피처 맵은 대상 참조 피처 맵으로 지칭될 수도 있다. 대상 패치 이미지가 쿼리 패치 이미지인 경우, 대상 쿼리 패치 이미지로 지칭되고, 대상 쿼리 패치 이미지에 대응하는 피처 맵은 대상 쿼리 피처 맵으로 지칭될 수도 있다.

[0058] 일 실시예에서, 초해상화 블록(112)은 쿼리 패치 이미지에 대응하는 피처 맵에 대해 상대적으로 적은 비트 수로 양자화를 수행할 수 있다. 예시적으로, 초해상화 블록(112)은 이미지 처리의 대상인 대상 패치 이미지가 쿼리 패치 이미지인 경우, 쿼리 패치 이미지에 대응하는 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 즉, 초해상화 블록(112)은 대상 쿼리 패치 이미지에 대응하는 대상 쿼리 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 실시예에서, 제2 비트 수는 제1 비트 수보다 적을 수 있다. 실시예에 따라, 초해상화 블록(112)은 대상 패치 이미지가 포함되는 클러스터 및 참조 패치 이미지인지 쿼리 패치 이미지인지 여부에 따라 상이한 비트 수로 양자화를 수행할 수도 있다.

[0059] 뉴럴 네트워크 프로세서(110)는 패치 이미지(PI)들에 이미지 처리 동작이 수행된 출력 패치 이미지들을 생성할 수 있다. 도 3에는 도시되지 않았으나, 실시예에 따라, 이미지 처리 장치(100)는 병합기를 포함할 수도 있다. 병합기는 출력 패치 이미지들을 병합하여 출력 이미지(ODT)로 생성할 수 도 있다. 다만, 이에 반드시 제한되는 것은 아니다.

[0060] 저해상도의 이미지를 고해상도의 이미지로 이미지 처리하기 위해 뉴럴 네트워크 연산이 수행되는 경우, 많은 연산량이 필요하다. 이미지 처리 장치(100)는 뉴럴 네트워크 연산 수행 시, 참조 패치 이미지에 대해 많은 비트 수로 양자화를 수행하고, 쿼리 패치 이미지에 대해 상대적으로 적은 비트 수로 양자화를 수행함으로써, 연산량을 감소시킬 수 있고, 이미지 처리 속도가 향상될 수 있다.

[0061] 도 4는 일 실시예에 따른 뉴럴 네트워크를 설명하기 위한 도면이다. 도 4의 뉴럴 네트워크(NN)는 도 3의 뉴럴 네트워크 프로세서(110)에 적용될 수 있다. 상술한 내용과 중복되는 내용은 생략한다.

[0062] 본 개시의 예시적 실시예에 따르면, 뉴럴 네트워크(NN)는 이미지 처리 동작을 수행할 수 있다. 뉴럴 네트워크(NN)는 저해상도의 이미지로부터 고해상도 이미지를 생성하는 초해상화 이미지 처리 동작을 수행하도록 훈련될 수 있다.

[0063] 뉴럴 네트워크(NN)는 클러스터링 블록(111), 잔차 블록(Residual Bolck), 컨볼루션 레이어(convolution layer)(c1, c2, c3), 선택기(113), 초해상화 블록(112), 에러 복원 블록(114), 업샘플링 레이어(upsampling layer)(us1, us2)를 포함할 수 있다. 뉴럴 네트워크(NN)는 클러스터링 블록(111), 잔차 블록(Residual Bolck), 컨볼루션 레이어(convolution layer)(c1, c2, c3), 선택기(113), 초해상화 블록(112), 에러 복원 블록(114), 업샘플링 레이어(upsampling layer)(us1, us2) 블록에 의해 패치 이미지(PI)의 피처 값을 연산할 수 있다. 예시적으로, 블록은 컨볼루션 레이어, 활성 레이어 등을 포함할 수 있다.

[0064] 본 개시의 예시적 실시예에 따르면, 클러스터링 블록(111)은 패치 이미지(PI)들을 수신할 수 있다. 클러스터링 블록(111)은 클러스터 수에 따라 유사한 패치 이미지들끼리 클러스터를 형성하도록 클러스터링할 수 있다. 클러

스터링 블록(111)은 특징 추출 블록(fb) 및 구별 블록(db)을 포함할 수 있다.

[0065] 특징 추출 블록(fb)은 패치 이미지(PI)들 각각의 특징을 추출할 수 있다. 특징 추출 블록(fb)은 특징 추출 과정을 통해 패치 이미지(PI)들 특징 정보를 생성할 수 있다. 예를 들어, 특징 추출 블록(fb)은 패치 이미지(PI)들 각각의 특징을 나타내는 특징 벡터를 생성할 수 있다. 특징 추출 블록(fb)은 복수의 레이어를 포함할 수 있다. 다만, 이에 반드시 제한되는 것은 아니며, 예시적으로, 특징 추출 블록(fb)은 LBF(local Binary Pattern), SIFT(Scale Invariant Feature Transform) 특징 추출기, vgg(visual geometry grouop) 16 등의 특징 추출기를 이용할 수 있다. 특징 추출 블록(fb)은 여러 개의 특징 추출기가 결합될 수도 있다. 또한, 실시예에 따라, 클러스터링 블록(111)에서 특징 추출 블록(fb)은 생략될 수도 있다.

[0066] 구별 블록(db)은 패치 이미지(PI)들 각각의 특징 정보를 이용하여 패치 이미지(PI)들을 클러스터링하고, 각 클러스터의 참조 패치 이미지 및 쿼리 패치 이미지로 구별할 수 있다. 예시적으로, 구별 블록(db)은 참조 패치 이미지인지 쿼리 패치 이미지인지 구별하는 구별 데이터를 생성할 수 있다. 구별 블록(db)은 특징 벡터를 이용하여 구별 데이터(dd)를 생성할 수 있다. 클러스터링 블록(db)에 대해 도 5를 참조하여 상세히 후술한다.

[0067] 구별 블록(db)은 복수의 레이어들을 포함할 수 있다. 구별 블록(db)은 소프트맥스(softmax) 함수 및 전치 소프트맥스(transpose softmax) 함수를 적용하여 구별 데이터(dd)를 생성할 수 있다. 다만, 이에 반드시 제한되는 것은 아니며, 구별 블록(db)은 패치 이미지(PI)들을 클러스터링하여 구별 데이터(dd)를 생성하기 위해 다양한 알고리즘을 사용할 수 있다. 예를 들어, 구별 블록(db)은 k- 평균 알고리즘(k-means algorithm), k-중앙자 알고리즘(k-medoids algorithm), dbscan(Density-Based Spatial Clustering of Applications with Noise)등의 알고리즘을 사용할 수 있다.

[0068] 패치 이미지(PI)들 각각은 컨볼루션 레이어(c1)로 입력될 수 있다. 예시적으로, 이미지 처리 장치의 이미지 처리 대상인 대상 패치 이미지(tPI)는 컨볼루션 레이어(c1)로 입력될 수 있다. 컨볼루션 레이어(c1)는 대상 패치 이미지(tPI)에 대해 컨볼루션 연산을 수행하여 대상 피처 맵(FMi)을 생성할 수 있다.

[0069] 대상 피처 맵(FMi)은 잔차 블록(ra1)으로 입력될 수 있다. 잔차 블록(ra1)은 대상 피처 맵(FMi)에 대해 잔차(residual) 연산을 수행하고 대상 피처 맵(FM0)을 생성할 수 있다. 잔차 블록(ra1)은 대상 피처 맵(FMi)에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다.

[0070] 선택기(113)는 구별 데이터(dd) 및 대상 피처 맵(FM0)을 수신할 수 있다. 선택기(113)는 구별 데이터(dd)에 기초하여 대상 피처 맵(FM0)의 피처 맵 유형을 판단할 수 있다. 선택기(113)는 대상 피처 맵(FM0)이 참조 피처 맵인지 쿼리 피처 맵인지 판단할 수 있다. 참조 피처 맵은 참조 패치 이미지에 대응하는 피처 맵을 의미하고, 쿼리 피처 맵은 쿼리 피처 이미지에 대응하는 피처 맵을 의미할 수 있다.

[0071] 구별 데이터(dd)는 어떤 클러스터의 참조 패치 이미지인지, 쿼리 패치 이미지인지에 대한 정보를 포함하므로, 선택기(113)는 구별 데이터(dd)를 이용하여 대상 피처 맵(FM0)이 어떤 클러스터의 참조 피처 맵인지 및 쿼리 피처 맵인지 여부를 판단할 수 있다.

[0072] 선택기(113)는 대상 피처 맵(FM0)이 참조 피처 맵에 해당하는 경우, 대상 피처 맵(FR0)으로 출력할 수 있다. 선택기(113)는 대상 피처 맵(FM0)이 쿼리 피처 맵에 해당하는 경우, 대상 피처 맵(FQ0)으로 출력할 수 있다. 선택기(113)대상 피처 맵(FM0)의 피처 맵 유형을 판단하여 초해상화 블록(112)으로 대상 피처 맵(FR0, FQ0)을 출력할 수 있다.

[0073] 초해상화 블록(112)은 대상 패치 이미지(tPI)에 대응하는 대상 피처 맵(FR0, FQ0)을 수신할 수 있다. 초해상화 블록(112)은 대상 피처 맵(FR0, FQ0)에 대해 이미지 처리를 위한 다양한 연산을 수행할 수 있다. 초해상화 블록(112)은 대상 피처 맵(FR0, FQ0)의 피처 맵 유형에 따라 서로 상이한 비트로 대상 피처 맵(FR0, FQ0)에 대해 양자화를 수행할 수 있다.

[0074] 일 실시예에서, 초해상화 블록(112)은 참조 피처 맵인 대상 피처 맵(FR0)을 수신하는 경우, 대상 피처 맵(FR0)에 대해 상대적으로 많은 비트 수로 양자화를 수행할 수 있다. 예시적으로, 초해상화 블록(112)은 대상 피처 맵(FR0)에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 초해상화 블록(112)은 대상 피처 맵(FR0)에 대해 양자화 동작 및 다른 이미지 처리 동작을 수행하고 대상 피처 맵(FRn)을 출력할 수 있다. 대상 피처 맵(FRn)은 업샘플링 레이어(us1) 및 컨볼루션 레이어(c2)를 순차적으로 거쳐 출력 패치 이미지(A1)로 생성될 수 있다.

[0075] 일 실시예에서, 초해상화 블록(112)은 쿼리 피처 맵인 대상 피처 맵(FQ0)을 수신하는 경우, 대상 피처 맵(FQ0)

에 대해 상대적으로 적은 비트 수로 양자화를 수행할 수 있다. 예시적으로, 초해상화 블록(112)은 대상 피처 맵(FQ0)에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 제2 비트 수는 제1 비트 수보다 적을 수 있다. 초해상화 블록(112)은 대상 피처 맵(FQ0)에 대해 양자화 동작 및 다른 이미지 처리 동작을 수행하고 대상 피처 맵(FQn-1)을 출력할 수 있다. 쿼리 패치 이미지에 대응하는 피처 맵은 상대적으로 적은 비트 수로 양자화되므로, 양자화 에러가 발생할 수 있다. 쿼리 패치 이미지에 대응하는 피처 맵에 대해 양자화를 수행 시, 발생되는 양자화 에러를 복원할 필요가 있다.

[0076] 대상 피처 맵(FQ0)은 상대적으로 적은 비트 수로 양자화되므로, 양자화 에러가 발생할 수 있다. 에러 복원 블록(114)은 대상 피처 맵(FQn-1)에 대해 양자화 에러를 복원하고, 대상 피처 맵(FQn)을 출력할 수 있다. 일 실시예에서, 에러 복원 블록(114)은 보조 피처 맵에 기초하여 대상 피처 맵(FQn-1)의 양자화 에러를 복원할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵을 의미할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵에 대해 양자화가 수행된 피처 맵일 수 있다. 대상 클러스터는 대상 패치 이미지가 포함된 클러스터일 수 있다.

[0077] 대상 피처 맵(FQn)은 업샘플링 레이어(us2) 및 컨볼루션 레이어(c3)를 순차적으로 거쳐 출력 패치 이미지(A2)로 생성될 수 있다. 도 4에는 뉴럴 네트워크(NN)에 하나의 에러 복원 블록(114)이 포함되는 것으로 도시되었으나, 이에 반드시 제한되는 것은 아니며, 2개 이상의 에러 복원 블록(114)이 포함될 수도 있다. 에러 복원 블록(114)에 대해 도 9a 및 도 9b에서 상세히 후술한다.

[0078] 도 5는 본 개시의 예시적인 실시예에 따른 클러스터링 블록을 설명하기 위한 도면이다. 상술한 내용과 중복되는 내용은 생략한다.

[0079] 도 5를 참조하면, 클러스터링 블록(111)은 복수의 레이어들을 포함할 수 있다. 클러스터링 블록(111)은 특징 추출 블록(fb) 및 구별 블록(db)을 포함할 수 있다.

[0080] 클러스터링 블록(111)은 패치 이미지(PI)들을 수신할 수 있다. 예시적으로, 클러스터링 블록(111)은 N(N은 양수)개의 패치 이미지(PI)들을 수신할 수 있다. 특징 추출 블록(fb)은 특징 추출 과정을 통해 패치 이미지(PI)들 특징 정보를 생성할 수 있다. 예를 들어, 특징 추출 블록(fb)은 패치 이미지(PI)들 각각의 특징을 나타내는 특징 벡터(V)를 생성할 수 있다. 특징 추출 블록(fb)은 복수의 레이어를 포함할 수 있다.

[0081] 도 5를 참조하면, 특징 추출 블록(fb)은 컨볼루션 레이어(c4, c5, c6), 풀링 레이어(p1), 및 풀리 커넥티드 레이어(fc1)를 포함할 수 있다. 특징 추출 블록(fb)은 패치 이미지(PI)들 각각에 대해 컨볼루션 레이어(c4), 컨볼루션 레이어(c5), 컨볼루션 레이어(c6), 풀링 레이어(p1), 및 풀리 커넥티드 레이어(fc1) 각각에 대응하는 연산을 수행하여 패치 이미지(PI)들 각각의 특징을 나타내는 특징 벡터(V)를 생성할 수 있다.

[0082] 구별 블록(db)은 패치 이미지(PI)들 각각의 특징 벡터(V)를 이용하여 패치 이미지(PI)들을 클러스터링하고, 각 클러스터의 참조 패치 이미지 및 쿼리 패치 이미지로 구별할 수 있다. 구별 블록(db)은 복수의 레이어들을 포함할 수 있다. 구별 블록(db)은 소프트맥스 레이어(s), 전치 소프트 맥스 레이어(s'), 곱셈기(multiplier), 및 아그맥스(argmax) 레이어(aml)를 포함할 수 있다.

[0083] 소프트맥스 레이어(s)는 패치 이미지(PI)들 각각에 대응하는 특징 벡터(V)에 소프트맥스 함수 연산을 적용하여 제1 분류 데이터(Adt)를 생성할 수 있다.

[0084] 전치 소프트맥스 레이어(s')는 패치 이미지(PI)들 각각에 대응하는 특징 벡터(V)에 전치 소프트맥스 함수 연산을 적용하여 제2 분류 데이터(Adt')를 생성할 수 있다. 전치 소프트맥스 함수는 특징 벡터(V)의 전치 행렬을 입력으로 하여 소프트맥스 함수를 적용하는 것을 의미할 수 있다.

[0085] 제1 분류 데이터(Adt) 및 제2 분류 데이터(Adt')를 곱한 곱셈 값(M)은 아그맥스 레이어(aml)의 입력일 수 있다. 아그맥스 레이어(aml)는 곱셈 값(M)에 아그맥스 함수 연산을 적용하여 구별 데이터(dd)를 생성할 수 있다.

[0086] 도 6은 본 개시의 예시적인 실시예에 따른 구별 블록의 동작을 설명하기 위한 도면이다. 도 5에서 상술한 내용과 중복되는 내용은 생략한다. 도 6의 C는 클러스터의 수를 의미하고, N는 패치 이미지의 개수를 나타낼 수 있다.

[0087] 도 5 및 도 6을 참조하면, 패치 이미지의 수는 4개이고, 클러스터의 수는 3개일수 있다. 다만, 도 6은 패치 이미지의 수가 4개인 예시를 나타내는 것으로, 패치 이미지의 수 및 클러스터의 수는 이에 반드시 제한되는 것은 아니며, 다양할 수 있다. 클러스터의 수는 기 설정되거나 학습에 따라 정해질 수 있다. 구별 블록(db)은 패치 이미지(PI)들 각각에 대응하는 특징 벡터(V)에 소프트맥스 함수 연산을 적용하여 제1 분류 데이터(Adt)를 생성

할 수 있다. 제1 분류 데이터(Adt)는 행렬로 나타낼 수 있다.

[0088] 구별 블록(db)은 패치 이미지(PI)들 각각에 대응하는 특징 벡터(V)에 전치 소프트맥스 함수 연산을 적용하여 제2 분류 데이터(Adt')를 생성할 수 있다. 제2 분류 데이터(Adt')는 행렬로 나타낼 수 있다. 구별 블록(db)은 제1 분류 데이터(Adt) 및 제2 분류 데이터(Adt')를 곱한 곱셈 값(M)에 아그맥스 함수 연산을 적용하여 구별 데이터(dd)를 생성할 수 있다.

[0089] 구별 데이터(dd)를 참조하면, 제1 클러스터에는 제1 패치 이미지, 제2 패치 이미지, 및 제3 패치 이미지가 포함될 수 있다. 제1 클러스터의 참조 패치 이미지는 제2 패치 이미지일 수 있다. 제1 클러스터의 쿼리 패치 이미지는 제1 패치 이미지 및 제3 패치 이미지일 수 있다. 제2 클러스터에는 제4 패치 이미지가 포함될 수 있다. 제2 클러스터의 참조 패치 이미지는 제4 패치 이미지이고, 제2 클러스터의 쿼리 패치 이미지는 없을 수 있다. 제3 클러스터에 포함되는 패치 이미지는 없을 수 있다.

[0090] 도 7은 본 개시의 예시적인 실시예에 따른 초해상화 블록을 나타내는 도면이다. 도 7의 초해상화 블록(112)은 도 4의 초해상화 블록(112)에 대응되므로, 중복되는 내용은 생략한다.

[0091] 초해상화 블록(112)은 대상 패치 이미지에 대응하는 대상 피처 맵(FR0, FQ0)을 수신할 수 있다. 초해상화 블록(112)은 복수의 잔차 블록들을 포함할 수 있다. 초해상화 블록(112)은 적어도 하나의 제1 잔차 블록(r1) 및 적어도 하나의 제2 잔차 블록(r2)을 포함할 수 있다.

[0092] 일 실시예에서, 초해상화 블록(112)에 포함된 적어도 하나의 제2 잔차 블록(r2)의 수는 초해상화 블록(112)에 포함된 적어도 하나의 제1 잔차 블록(r1)의 수보다 적을 수 있다. 예를 들어, 초해상화 블록(112)은 n(n은 양수)개의 제1 잔차 블록(r1)을 포함하고, n-1개의 제2 잔차 블록(r2)을 포함할 수 있다. 다만, 이에 반드시 제한되는 것은 아니다. 예시적으로, 초해상화 블록(112)에 포함되는 적어도 하나의 제1 잔차 블록(r1)의 수는, 초해상화 블록(112)에 포함되는 적어도 하나의 제2 잔차 블록(r2)의 수 및 에러 복원 블록(예를 들어, 도 4의 에러 복원 블록(114))의 수의 합과 동일할 수 있다.

[0093] 일 실시예에서, 적어도 하나의 제1 잔차 블록(r1)은 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 적어도 하나의 제1 잔차 블록(r1) 각각은 적어도 하나의 제1 잔차 블록(r1) 각각으로 입력되는 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 참조 피처 맵이 제1 잔차 블록(r1_1)으로 입력될 수 있다. 대상 피처 맵(FR0)이 참조 피처 맵인 경우, 적어도 하나의 제1 잔차 블록(r1)들 각은 입력되는 참조 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다.

[0094] 적어도 하나의 제1 잔차 블록(r1) 각각은, 제1 잔차 블록(r1) 각각으로 입력되는 피처 맵 각각에 대한 참조 액티베이션들 및 참조 웨이트들의 양자화를 수행할 수 있다. 적어도 하나의 제1 잔차 블록(r1) 각각은, 제1 비트 수에 기초하여 참조 액티베이션들 및 참조 웨이트들에 대한 양자화를 수행할 수 있다. 예를 들어, 적어도 하나의 제1 잔차 블록(r1)은 대상 피처 맵(FR0)에 대해 8비트 수에 기초하여 양자화를 수행할 수 있다. 다만, 이에 반드시 제한되는 것은 아니며, 대상 피처 맵(FR0)은 다양한 비트수로 양자화될 수 있다.

[0095] 예시적으로, 제1 잔차 블록(r1_1)은 대상 피처 맵(FR0)에 대해 제1 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FR1)을 출력할 수 있다. 대상 피처 맵(FR1)은 제1 잔차 블록(r1_2)의 입력일 수 있고, 제1 잔차 블록(r1_2)은 대상 피처 맵(FR1)에 대해 제1 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FR2)을 출력할 수 있다. 대상 피처 맵(FRn-1)은 제1 잔차 블록(r1_n)의 입력일 수 있고, 제1 잔차 블록(r1_n)은 대상 피처 맵(FRn-1)에 대해 제1 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FRn)을 출력할 수 있다. 대상 피처 맵(FRn)은 업샘플링 레이어(예를 들어, 도 4의 업샘플링 레이어(us1))로 입력될 수 있다.

[0096] 일 실시예에서, 적어도 하나의 제2 잔차 블록(r2)은 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 적어도 하나의 제2 잔차 블록(r2) 각각은 적어도 하나의 제2 잔차 블록(r2) 각각으로 입력되는 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 제2 비트 수는 제1 비트 수보다 적을 수 있다. 쿼리 피처 맵이 제2 잔차 블록(r1_2)으로 입력될 수 있다. 대상 피처 맵(FQ0)이 쿼리 피처 맵인 경우, 적어도 하나의 제2 잔차 블록(r2)들 각각은 입력되는 쿼리 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다.

[0097] 예시적으로, 제2 잔차 블록(r2_1)은 대상 피처 맵(FQ0)에 대해 제2 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FQ1)을 출력할 수 있다. 대상 피처 맵(FQ1)은 제2 잔차 블록(r2_2)의 입력일 수 있고, 제2 잔차 블록(r2_2)은 대상 피처 맵(FQ1)에 대해 제2 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FQ2)을 출력할 수 있다. 대상 피처 맵(FQn-2)은 제2 잔차 블록(r2_n-1)의 입력일 수 있고, 제2 잔차 블록(r2_n-1)은 대상 피처 맵

(FQn-2)에 대해 제2 비트 수에 기초하여 양자화를 수행하고 대상 피처 맵(FQn-1)을 출력할 수 있다. 대상 피처 맵(FQn-1)은 에러 복원 블록(예를 들어, 도 4의 에러 복원 블록(114))으로 입력될 수 있다.

[0098] 적어도 하나의 제2 잔차 블록(r2) 각각은, 적어도 하나의 제2 잔차 블록(r2) 각각으로 입력되는 피처 맵 각각에 대한 쿼리 액티베이션들 및 쿼리 웨이트들의 양자화를 수행할 수 있다. 적어도 하나의 제2 잔차 블록(r2)은 쿼리 액티베이션들 및 쿼리 웨이트들 각각에 대해 상이한 비트 수로 양자화를 수행할 수 있다.

[0099] 일 실시예에서, 적어도 하나의 제2 잔차 블록(r2) 각각은, 제2 비트 수에 기초하여 쿼리 액티베이션들에 대한 양자화를 수행할 수 있다. 적어도 하나의 제2 잔차 블록(r2) 각각은, 제3 비트 수에 기초하여 쿼리 가중치들에 대한 양자화를 수행할 수 있다. 제3 비트 수는 제1 비트 수보다 적고, 제2 비트 수보다 많을 수 있다.

[0100] 예시적으로, 제3 비트 수는 제1 비트 수와 제2 비트 수의 중간 값일 수있다. 예를 들어, 제2 잔차 블록(r2_1)은 대상 피처 맵(FQ0)의 쿼리 액티베이션들에 대해 4비트 수에 기초하여 양자화를 수행할 수 있다. 제2 잔차 블록(r2_1)은 대상 피처 맵(FQ0)의 쿼리 가중치들에 대해 6비트 수에 기초하여 양자화를 수행할 수 있다. 다만, 이에 반드시 제한되는 것은 아니며, 쿼리 액티베이션들 및 쿼리 가중치들은 다양한 비트수로 양자화될 수 있다.

[0101] 이미지 처리 장치는 뉴럴 네트워크 연산 수행 시, 참조 패치 이미지에 대응하는 피처 맵에 대해 많은 비트 수로 양자화를 수행하고, 쿼리 패치 이미지에 대응하는 피처 맵에 대해 상대적으로 적은 비트 수로 양자화를 수행함으로써, 뉴럴 네트워크의 연산량을 감소시킬 수 있다.

[0102] 도 8은 본 개시의 예시적인 실시예에 따른 잔차 블록을 설명하기 위한 도면이다. 도 8의 잔차 블록은 제1 잔차 블록(예를 들어, 도 7의 제1 잔차 블록) 및 제2 잔차 블록(예를 들어, 도 7의 제2 잔차 블록)에 적용될 수 있다.

[0103] 잔차 블록은 적어도 하나의 양자화 컨볼루션 레이어(quantization convolution layer) 및 적어도 하나의 활성 레이어(activation layer)를 포함할 수 있다. 예시적으로, 잔차 블록은 2개의 양자화 컨볼루션 레이어 및 1개의 활성 레이어를 포함할 수 있으나, 양자화 컨볼루션 레이어 및 활성 레이어의 수는 이에 반드시 제한되는 것은 아니다.

[0104] 양자화 컨볼루션 레이어는 입력되는 피처 맵에 대해 양자화 컨볼루션 연산을 수행할 수 있다. 양자화 컨볼루션 레이어는 적어도 하나의 양자화 레이어, 적어도 하나의 컨볼루션 레이어, 및 적어도 하나의 역양자화 레이어(dequantization layer)를 포함할 수 있다. 예시적으로, 양자화 컨볼루션 레이어는 1개의 양자화 레이어, 1개의 컨볼루션 레이어, 및 1개의 역양자화 레이어를 포함할 수 있으나, 이에 반드시 제한되는 것은 아니다. 양자화 레이어는 입력되는 피처 맵의 파라미터들에 대해 양자화 연산을 수행할 수 있다. 역양자화 레이어는 입력되는 피처 맵의 파라미터들에 대해 역양자화 연산을 수행할 수 있다.

[0105] 활성 레이어는 활성 레이어로 입력되는 피처 맵에 대해 활성 함수 연산을 수행할 수 있다. 예시적으로, 활성 레이어는 ReLU(Rectified Linear Unit) 함수에 의한 연산을 수행할 수 있다. 다만, 이에 반드시 제한되는 것은 아니며, 활성 레이어는 시그모이드(sigmoid) 함수 등 다양한 활성 함수에 의한 연산을 수행할 수 있다.

[0106] 잔차 블록의 입력 각각에 대해 양자화 컨볼루션 연산 및 활성 함수 연산이 수행될 수 있다. 잔차 블록은 잔차 블록으로 입력되는 피처 맵에 대해 양자화 컨볼루션 연산 및 활성 함수 연산을 수행하여 피처 맵에 대해 양자화를 수행할 수 있다. 예시적으로, 잔차 블록으로 입력되는 피처 맵의 피처 맵 유형에 따라, 양자화를 수행하는 비트 수가 상이할 수 있다.

[0107] 도 9a는 본 개시의 예시적인 실시예에 따른 에러 복원 블록을 설명하기 위한 도면이다. 도 9a의 초해상화 블록(112) 및 에러 복원 블록(114)에 대해 상술한 내용과 중복되는 내용은 생략한다.

[0108] 도 9a를 참조하면, 뉴럴 네트워크(NN)는 초해상화 블록(112) 및 에러 복원 블록(114)을 포함할 수 있다. 에러 복원 블록(114)은 쿼리 이미지에 대해 양자화 에러를 복원할 수 있다. 에러 복원 블록(114)은 쿼리 이미지에 대응하는 쿼리 피처 맵의 양자화 에러를 복원할 수 있다. 에러 복원 블록(114)은 대상 클러스터에 포함된 쿼리 패치 이미지들 중 이미지 처리 대상인 대상 쿼리 패치 이미지에 대해 양자화 에러를 복원할 수 있다.

[0109] 초해상화 블록은 쿼리 피처 맵인 대상 피처 맵(FQ0)에 대해 양자화를 수행하여 대상 피처 맵(FQn-1)을 출력하고, 에러 복원 블록(114)은 쿼리 피처 맵인 대상 피처 맵(FQn-1)의 양자화 에러를 복원할 수 있다. 다시 말해, 에러 복원 블록(114)은 초해상화 블록(112)에서 양자화가 수행된 대상 피처 맵(FQn-1)의 양자화 에러를 복원하고, 대상 피처 맵(FQn)을 출력할 수 있다.

[0110] 일 실시예에서, 에러 복원 블록(114)은 보조 피처 맵에 기초하여 대상 피처 맵(FQn-1)의 양자화 에러를 복원할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵을 의미할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵에 대해 양자화가 수행된 피처 맵일 수 있다. 보조 피처 맵은 제1 잔차 블록들로부터 양자화된 피처 맵이므로, 많은 비트 수로 양자화될 수 있다.

[0111] 예시적으로, 보조 피처 맵은 에러 복원 블록(114)로 입력되는 쿼리 피처 맵을 출력하는 제2 잔차 블록과 대응하는 블록으로부터 출력되는 피처 맵일 수 있다. 보조 피처 맵은 제2 잔차 블록과 대응하는 제1 잔차 블록으로부터 출력되는 피처 맵일 수 있다. 예를 들어, 제2 잔차 블록(r2n-1)은 쿼리 피처 맵인 대상 피처 맵(FQn-1)을 출력하고, 제2 잔차 블록(r2n-1)은 제1 잔차 블록(r1n-1)과 대응할 수 있다. 참조 피처 맵(FRn-1)이 보조 피처 맵으로서 에러 복원 블록(114)에 제공될 수 있다. 에러 복원 블록(114)은 참조 피처 맵(FRn-1)에 기초하여 대상 피처 맵(FQn-1)의 양자화 에러를 복원할 수 있다.

[0112] 도 9a에는 에러 복원 블록(114)이 1개인 것으로 도시되어 있으나, 이에 반드시 제한되는 것은 아니며, 이미지 처리 장치에 포함되는 에러 복원 블록(114)의 수는 1개 이상일 수도 있다. 많은 비트 수로 양자화된 보조 피처 맵을 이용하여 쿼리 피처 맵에 대해 양자화 에러가 복원됨으로써, 쿼리 피처 맵의 양자화 에러가 효율적이고 높은 정확도로 극복될 수 있다.

[0113] 도 9b는 본 개시의 예시적인 실시예에 따른 에러 복원 블록이 복수 개인 경우를 설명하기 위한 도면이다. 도 9a와 비교하면, 도 9b의 뉴럴 네트워크(NN)는 2개의 에러 복원 블록을 포함할 수 있다. 상술한 내용과 중복되는 내용은 생략한다.

[0114] 도 9b를 참조하면, 뉴럴 네트워크(NN)는 제1 에러 복원 블록(114_1) 및 제2 에러 복원 블록(114_2)을 포함할 수 있다. 제1 에러 복원 블록(114_1)은 대상 피처 맵(FQn-1)의 양자화 에러를 복원하고, 대상 피처 맵(FQn)을 출력할 수 있다. 제2 에러 복원 블록(114_2)은 대상 피처 맵(FQn)의 양자화 에러를 복원하고, 대상 피처 맵(FQn+1)을 출력할 수 있다. 제2 에러 복원 블록(114_2)은 제1 에러 복원 블록(114_1)으로부터 출력된 대상 피처 맵(FQn)의 양자화 에러를 복원할 수 있다. 대상 피처 맵(FQn+1)은 업샘플링 레이어및 컨볼루션 레이어를 순차적으로 거쳐 출력 패치 이미지로 생성될 수 있다.

[0115] 실시예에서, 초해상화 블록(112)에 포함되는 적어도 하나의 제1 잔차 블록(r1)의 수는, 초해상화 블록(112)에 포함되는 적어도 하나의 제2 잔차 블록(r2)의 수 및 에러 복원 블록(114_1, 114_2)의 수의 합과 동일할 수 있다. 예를 들어, 초해상화 블록(112)에 포함되는 제1 잔차 블록(r1)의 수는 n+1개일 수 있다. 초해상화 블록(112)에 포함되는 제2 잔차 블록(r2)의 수는 n-1개이고, 에러 복원 블록(114_1, 114_2)의 수는 2개 이므로, 이들의 합은 n+1일 수 있다.

[0116] 일 실시예에서, 에러 복원 블록(114_1, 114_2)은 보조 피처 맵에 기초하여 대상 피처 맵의 양자화 에러를 복원할 수 있다. 예시적으로, 보조 피처 맵은 제1 에러 복원 블록(114_1)으로 입력되는 쿼리 피처 맵(FQn-1)을 출력하는 제2 잔차 블록(r2_n-1)과 대응하는 제1 잔차 블록(r1_n-1)으로부터 출력되는 피처 맵일 수 있다. 제2 잔차 블록(r2n-1)은 제1 잔차 블록(r1n-1)과 대응할 수 있고, 참조 피처 맵(FRn-1)이 보조 피처 맵으로서 제1 에러 복원 블록(114_1)에 제공될 수 있다. 제1 에러 복원 블록(114_1)은 참조 피처 맵(FRn-1)에 기초하여 대상 피처 맵(FQn-1)의 양자화 에러를 복원하고, 대상 피처 맵(FQn)을 출력할 수 있다.

[0117] 예시적으로, 보조 피처 맵은 제2 에러 복원 블록(114_2)으로 입력되는 쿼리 피처 맵(FQn)을 출력하는 에러 복원 블록(114_1, 114_2)과 대응하는 제1 잔차 블록(r1_n)으로부터 출력되는 피처 맵일 수 있다. 제1 에러 복원 블록(114_1)은 제1 잔차 블록(r1_n)과 대응할 수 있고, 참조 피처 맵(FRn)이 보조 피처 맵으로서 제2 에러 복원 블록(114_2)으로 제공될 수 있다. 제2 에러 복원 블록(114_2)은 참조 피처 맵(FRn)에 기초하여 대상 피처 맵(FQn)의 양자화 에러를 복원하고, 대상 피처 맵(FQn+1)을 출력할 수 있다.

[0118] 도 10은 본 개시의 예시적 실시예에 따른 에러 복원 블록의 구조를 설명하기 위한 도면이다. 도 10의 에러 복원 블록(114)은 도 9a 및 도 9b에서 설명한 에러 복원 블록(114)에 대응되므로, 중복되는 내용은 생략한다. 도 10에서는, 설명의 편의를 위해 에러 복원 블록(114)으로 대상 피처 맵(FQn-1)이 입력되고, 보조 피처 맵(FRn-1)이 제공되는 경우를 가정하여 설명한다. 대상 피처 맵(FQn-1)은 쿼리 피처 맵이고, 보조 피처 맵(FRn-1)은 참조 피처 맵일 수 있다.

[0119] 도 10을 참고하면, 에러 복원 블록(114)은 복수의 레이어들, 감산기(subtracter), 및 합산기(adder)를 포함할 수 있다. 복수의 레이어들은, 적어도 하나의 워핑(warping) 연산 레이어(wl), 적어도 하나의 양자화 컨볼루션 레이어(ql1, ql2), 및 적어도 하나의 활성 레이어(al)를 포함할 수 있다. 예시적으로, 에러 복원 블록(114)은 1

개의 워핑 연산 레이어(wl), 2개의 양자화 컨볼루션 레이어(ql1, ql2), 및 1개의 활성 레이어(al)를 포함할 수 있으나, 워핑 연산 레이어(wl), 컨볼루션 레이어(ql1, ql2), 및 활성 레이어(al)의 수는 이에 반드시 제한되는 것은 아니다.

[0120] 워핑 연산 레이어(wl)는 보조 피처 맵(FRn-1)에 워핑 연산을 적용하여 워핑된 보조 피처 맵(sFr)을 생성할 수 있다. 워핑 연산 레이어(wl)는 보조 피처 맵(FRrn-1)을 대상 피처 맵(FQn-1)에 워핑할 수 있다. 워핑 연산은 두 이미지의 중요한 특징을 좀 더 가깝게 정렬하는 것으로, 워핑 연산 레이어(wl)는 보조 피처 맵(FRn-1)을 대상 피처 맵(FQn-1)에 맞게 정렬할 수 있다. 워핑 연산 레이어(wl)는 다양한 워핑 연산 기법을 이용하여 워핑 연산을 수행할 수 있다.

[0121] 에러 복원 블록(114)은 워핑된 보조 피처 맵(sFR) 및 대상 피처 맵(FQn-1)을 감산 연산하여 감산 결과(ssr)를 생성할 수 있다. 워핑된 보조 피처 맵(sFR) 및 대상 피처 맵(FQn-1)이 감산 연산되면, 대상 피처 맵(FQn-1)의 양자화 에러가 남을 수 있다. 감산 결과(ssr)는 대상 피처 맵(FQn-1)의 양자화 에러를 나타낼 수 있다.

[0122] 에러 복원 블록(114)은 감산 결과(ssr)에 양자화 및 활성 함수 연산을 수행하여 복원 쿼리 피처 맵(rFM)을 생성할 수 있다. 에러 복원 블록(114)은 감산 결과(ssr)에 대해 양자화 컨볼루션 연산 및 활성 함수 연산을 수행하여 복원 쿼리 피처 맵(rFM)을 생성할 수 있다. 감산 결과(ssr)는 제1 양자화 컨볼루션 레이어(ql1), 활성 레이어(al), 및 제2 양자화 컨볼루션 레이어(ql2)를 통해 복원 쿼리 피처 맵(rFM)으로 생성될 수 있다.

[0123] 에러 복원 블록(114)은 감산 결과(ssr)에 대해 양자화를 수행하여 복원 쿼리 피처 맵(rFM)을 생성할 수 있다. 에러 복원 블록(114)은 제1 양자화 컨볼루션 레이어(ql1) 및 제2 양자화 컨볼루션 레이어(ql2)에서 양자화를 수행할 수 있다. 에러 복원 블록(114)은 상대적으로 많은 비트 수에 기초하여 양자화를 수행할 수 있다. 일 실시예에서, 에러 복원 블록(114)은 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 예시적으로, 에러 복원 블록(114)은 감산 결과(ssr)에 대해 제1 양자화 컨볼루션 레이어(ql1) 및 제2 양자화 컨볼루션 레이어(ql2)에서 제1 비트 수에 기초하여 양자화를 수행할 수 있다.

[0124] 에러 복원 블록(114)은 복원 쿼리 피처 맵(rFM) 및 대상 피처 맵(FQn-1)을 합산하는 합산 연산을 수행하고, 대상 피처 맵(FQn)을 생성할 수 있다. 대상 피처 맵(FQn)은 양자화 에러가 복원된 쿼리 피처 맵일 수 있다.

[0125] 에러 복원 블록(114)은 적은 비트 수로 양자화된 쿼리 피처 맵에 대해 많은 비트 수로 양자화함으로써, 쿼리 피처 맵의 양자화 에러를 복원할 수 있다. 따라서, 이미지 처리 장치로부터 출력되는 이미지의 품질이 향상될 수 있다.

[0126] 도 11은 본 개시의 예시적 실시 예에 따른 이미지 처리 방법을 나타내는 흐름도이다. 도 11은 이미지 처리 장치의 동작 방법을 나타내는 흐름도이다. 상세하게는, 도 11은 전술한 뉴럴 네트워크 프로세서의 이미지 처리 방법을 나타내는 흐름도이다.

[0127] 단계 S1110에서, 이미지 처리 장치는 복수의 패치 이미지들을 획득할 수 있다. 이미지 처리 장치는 이력 이미지가 분할된 복수의 패치 이미지들을 획득할 수 있다. 예시적으로, 패치 이미지들은 서로 동일한 크기일 수도 있고, 상이한 크기일 수도 있다.

[0128] 단계 S1120에서, 이미지 처리 장치는 복수의 패치 이미지들을 클러스터링하여 구별 데이터를 생성할 수 있다. 구별 데이터는 복수의 패치 이미지들이 클러스터링되어 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 데이터일 수 있다.

[0129] 이미지 처리 장치는 클러스터 수에 따라 유사한 패치 이미지들끼리 클러스터를 형성하도록 클러스터링할 수 있다. 예시적으로, 이미지 처리 장치는 제1 패치 이미지 및 제2 패치 이미지를 제1 클러스터로 클러스터링하고, 제3 패치 이미지 및 제4 패치 이미지를 제2 클러스터로 클러스터링할 수 있다. 다만, 이는 일 예시일 뿐 이에 반드시 제한되는 것은 아니다.

[0130] 이미지 처리 장치는 각 클러스터에 포함된 적어도 하나의 패치 이미지들을 참조 패치 이미지 및 쿼리 패치 이미지로 구별할 수 있다. 이미지 처리 장치는 각 클러스터의 특징을 가장 잘 나타내는 패치 이미지를 각 클러스터의 참조 패치 이미지로 구별할 수 있다. 예시적으로, 이미지 처리 장치는 제1 클러스터의 특징을 가장 잘 나타내는 제1 패치 이미지를 제1 클러스터의 참조 패치 이미지로 구별할 수 있다.

[0131] 이미지 처리 장치는 각 클러스터에 포함된 패치 이미지들 중 참조 패치 이미지가 아닌 패치 이미지를 쿼리 패치 이미지로 구별할 수 있다. 예를 들어, 이미지 처리 장치는 제1 클러스터에서 참조 패치 이미지가 아닌 제2 패치

이미지를 제1 클러스터의 쿼리 패치 이미지로 구별할 수 있다.

[0132] 단계 S1130에서, 이미지 처리 장치는 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성할 수 있다. 대상 패치 이미지는 이미지 처리 장치가 이미지 처리를 수행하는 대상인 패치 이미지를 의미할 수 있다. 예시적으로, 이미지 처리 장치는 대상 패치 이미지에 대해 컨볼루션 연산 및 잔차 연산을 수행하여 대상 피처 맵을 생성할 수 있다. 도 11에서, 단계 S1130은 단계 S1120에 후행하는 것으로 도시되어 있으나, 이에 반드시 제한되는 것은 아니다. 단계 S1130은 단계 S1120과 동시에 진행될 수도 있고, 단계 S1120에 선행하여 진행될 수도 있다.

[0133] 단계 S1140에서, 이미지 처리 장치는 구별 데이터에 기초하여 대상 피처 맵의 피처 맵 유형을 판단할 수 있다. 이미지 처리 장치는 대상 피처 맵이 참조 피처 맵인지 쿼리 피처 맵인지 판단할 수 있다. 참조 피처 맵은 참조 패치 이미지에 대응하는 피처 맵을 의미하고, 쿼리 피처 맵은 쿼리 피처 이미지에 대응하는 피처 맵을 의미할 수 있다.

[0134] 구별 데이터는 어떤 클러스터의 참조 패치 이미지인지, 쿼리 패치 이미지인지에 대한 정보를 포함하므로, 이미지 처리 장치는 구별 데이터를 이용하여 대상 피처 맵이 어떤 클러스터의 참조 피처 맵인지 및 쿼리 피처 맵인지 여부를 판단할 수 있다.

[0135] 단계 S1150에서, 이미지 처리 장치는 대상 피처 맵의 유형에 따라 서로 상이한 비트로 대상 피처 맵에 대해 양자화를 수행할 수 있다. 일 실시예에서, 이미지 처리 장치는 참조 피처 맵인 대상 피처 맵을 수신하는 경우, 대상 피처 맵에 대해 상대적으로 많은 비트 수로 양자화를 수행할 수 있다.

[0136] 일 실시예에서, 이미지 처리 장치는 쿼리 피처 맵인 대상 피처 맵을 수신하는 경우, 대상 피처 맵에 대해 상대적으로 적은 비트 수로 양자화를 수행할 수 있다. 이하에서는, 도 12를 참조하여 단계 S1150에 대해 설명한다.

[0137] 도 12는 본 개시의 예시적인 실시예에 따른 이미지 처리 장치가 양자화를 수행하는 방법을 나타내는 흐름도이다. 도 12의 단계들은 도 11의 단계 S1150에 포함될 수 있다.

[0138] 단계 S1210에서, 이미지 처리 장치는 대상 피처 맵의 피처 맵 유형이 참조 피처 맵인지 여부를 판단할 수 있다. 이미지 처리 장치는, 대상 피처 맵의 피처 맵 유형이 참조 피처 맵인 경우, 단계 S1220을 수행하고, 대상 피처 맵의 피처 맵 유형이 참조 피처 맵이 아닌 경우, 단계 S1230을 수행할 수 있다.

[0139] 단계 S1220에서, 이미지 처리 장치는 대상 피처 맵에 대해 제1 비트 수에 기초하여 양자화를 수행할 수 있다. 이미지 처리 장치는 대상 피처 맵에 대해 양자화 동작 및 다른 이미지 처리 동작을 수행할 수 있다. 예시적으로, 이미지 처리 장치는 대상 피처 맵의 피처 맵 유형이 참조 피처 맵인 경우, 8비트에 기초하여 양자화를 수행할 수 있다. 다만, 이에 반드시 제한되는 것은 아니다.

[0140] 단계 S1230에서, 이미지 처리 장치는 대상 피처 맵에 대해 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 이미지 처리 장치는, 대상 피처 맵의 피처 맵 유형이 참조 피처 맵이 아닌 경우, 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 이미지 처리 장치는, 대상 피처 맵의 피처 맵 유형이 쿼리 피처 맵인 경우, 제2 비트 수에 기초하여 양자화를 수행할 수 있다.

[0141] 일 실시예에서, 제2 비트 수는 제1 비트 수보다 적을 수 있다. 예시적으로, 이미지 처리 장치는 대상 피처 맵의 피처 맵 유형이 쿼리 피처 맵인 경우, 4비트에 기초하여 양자화를 수행할 수 있다. 다만, 이에 반드시 제한되는 것은 아니다.

[0142] 이미지 처리 장치는 쿼리 패치 이미지에 대해 양자화 에러를 복원하는 단계를 더 수행할 수 있다. 이미지 처리 장치는 보조 피처 맵에 기초하여 쿼리 피처 맵의 양자화 에러를 복원할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵을 의미할 수 있다. 보조 피처 맵은 대상 클러스터에 포함된 참조 패치 이미지에 대응하는 피처 맵에 대해 양자화가 수행된 피처 맵일 수 있다.

[0143] 도 13은 본발명의 일 실시예에 따른 이미지 처리 장치를 설명하기 위한 블록도이다. 도 13의 이미지 처리 장치(1000)는 도 1의 이미지 처리 장치(100)에 대응되므로, 중복되는 내용은 생략한다.

[0144] 도 13을 참조하면, 이미지 처리 장치(1000)는 메모리(1010) 및 프로세서(1020)를 포함할 수 있다. 메모리(1010)는 프로세서(1020)에서 실행되는 하나 이상의 인스트럭션들을 저장할 수 있다. 예를 들면, 메모리(1010)는 프로세서(1020)가 입력 이미지 데이터에 대해 이미지 처리 동작을 수행하는 인스트럭션을 포함할 수 있다.

[0145] 프로세서(1020)는 하나 이상의 인스트럭션을 실행함으로써 이미지 처리 동작을 수행할 수 있다. 뉴럴 네트워크

(예를 들어, 도 4의 뉴럴 네트워크(NN))는 프로세서(1020)에 적용될 수 있다. 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여 입력 이미지에 대한 양자화 동작을 수행하고, 출력 이미지를 생성할 수 있다. 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여 복수의 패치 이미지들을 획득할 수 있다. 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여 복수의 패치 이미지들을 클러스터링하여 각 클러스터에 포함된 적어도 하나의 패치 이미지들이 참조 패치 이미지에 해당하는 지 여부 및 쿼리 패치 이미지에 해당하는 지 여부를 구별하는 구별 데이터를 생성할 수 있다.

[0146] 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여 복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성할 수 있다. 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여 구별 데이터에 기초하여 상기 대상 피처 맵의 피처 맵 유형을 판단할 수 있다. 프로세서는 하나 이상의 인스트럭션을 실행하여 대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트로 대상 피처 맵에 대해 양자화를 수행할 수 있다.

[0147] 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여, 대상 피처 맵의 피처 맵 유형이 참조 피처 맵인 경우, 제1 비트 수에 기초하여 양자화를 수행하고, 대상 피처 맵의 피처 맵 유형이 쿼리 피처 맵인 경우 제2 비트 수에 기초하여 양자화를 수행할 수 있다. 프로세서(1020)는 하나 이상의 인스트럭션을 실행하여, 대상 클러스터에 포함된 참조 패치 이미지에 대응하고, 제1 비트 수에 기초하여 양자화가 수행된 보조 피처 맵을 이용하여, 제2 비트 수에 기초하여 양자화가 수행된 대상 피처 맵의 양자화 에러를 복원할 수 있다.

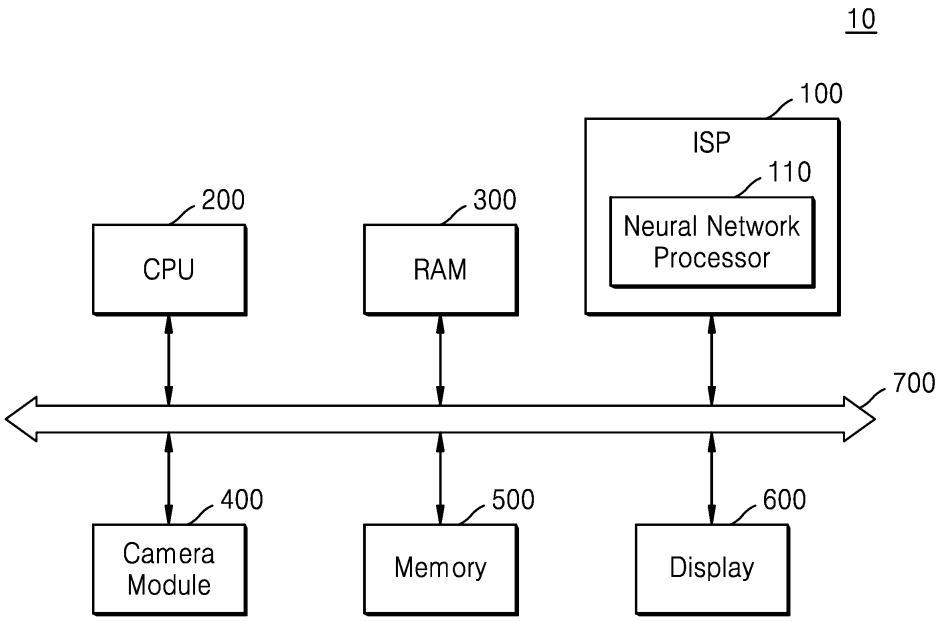
[0148] 메모리(1010)는 데이터를 저장하기 위한 저장 장소로서, 예를 들어, 다양한 알고리즘, 각종 프로그램 및 각종 데이터를 저장할 수 있다. 메모리(1010)는 하나 이상의 인스트럭션을 저장할 수 있다. 메모리(1010)는 휘발성 메모리(voltatile memory) 및 비휘발성 메모리(non-volatile memory) 중(among) 적어도 하나를 포함할 수 있다. 비휘발성 메모리는 ROM(Read Only Memory), PROM(Programmable ROM), EPROM(Electrically Programmable ROM), EEPROM(Electrically Erasable and Programmable ROM), 플래시 메모리, PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM) 등을 포함할 수 있다.

[0149] 휘발성 메모리는 DRAM(Dynamic RAM), SRAM(Static RAM), SDRAM(Synchronous DRAM), PRAM(Phase-change RAM), MRAM(Magnetic RAM), RRAM(Resistive RAM) 등을 포함할 수 있다. 또한 일 실시 예에 있어서, 메모리(1010)는 HDD(Hard Disk Drive), SSD(Solid State Drive), CF(Compact Flash), SD(Secure Digital), Micro-SD(Micro Secure Digital), Mini-SD(Mini Secure Digital), xD(extreme digital) 또는 Memory Stick 중 적어도 하나를 포함할 수도 있다. 일 실시 예에서, 메모리(1010)는 프로세서(1020)에 의해 실행되는 알고리즘들, 프로그램들 및 하나 이상의 인스트럭션들을 반영구적으로 또는 일시적으로 저장할 수 있다.

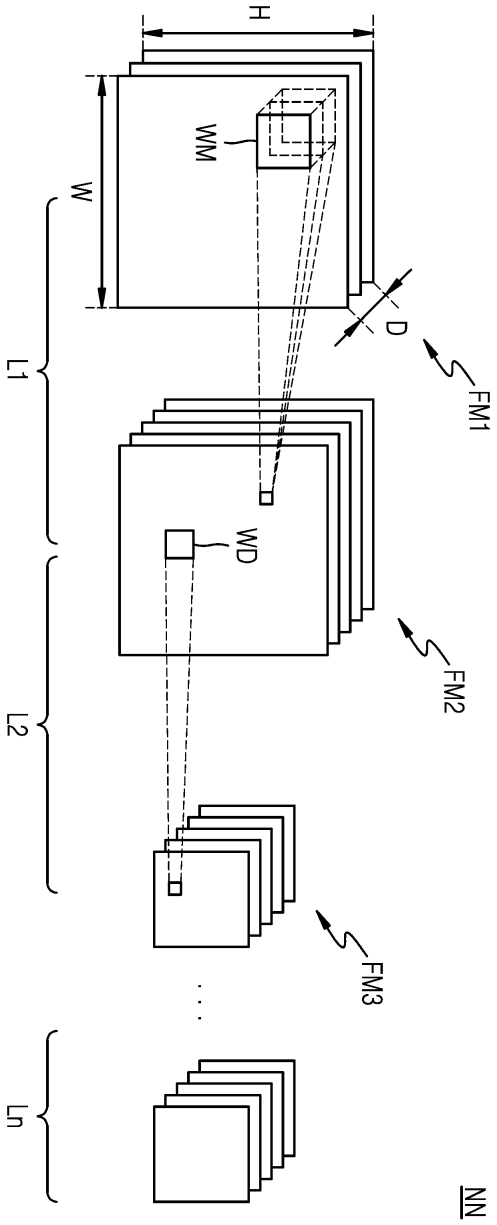
[0150] 이상에서와 같이 도면과 명세서에서 예시적인 실시 예들이 개시되었다. 본 명세서에서 특정한 용어를 사용하여 실시 예들을 설명되었으나, 이는 단지 본 개시의 기술적 사상을 설명하기 위한 목적에서 사용된 것이지 의미 한정이나 특허청구범위에 기재된 본 개시의 범위를 제한하기 위하여 사용된 것은 아니다. 그러므로 본 기술분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 개시의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

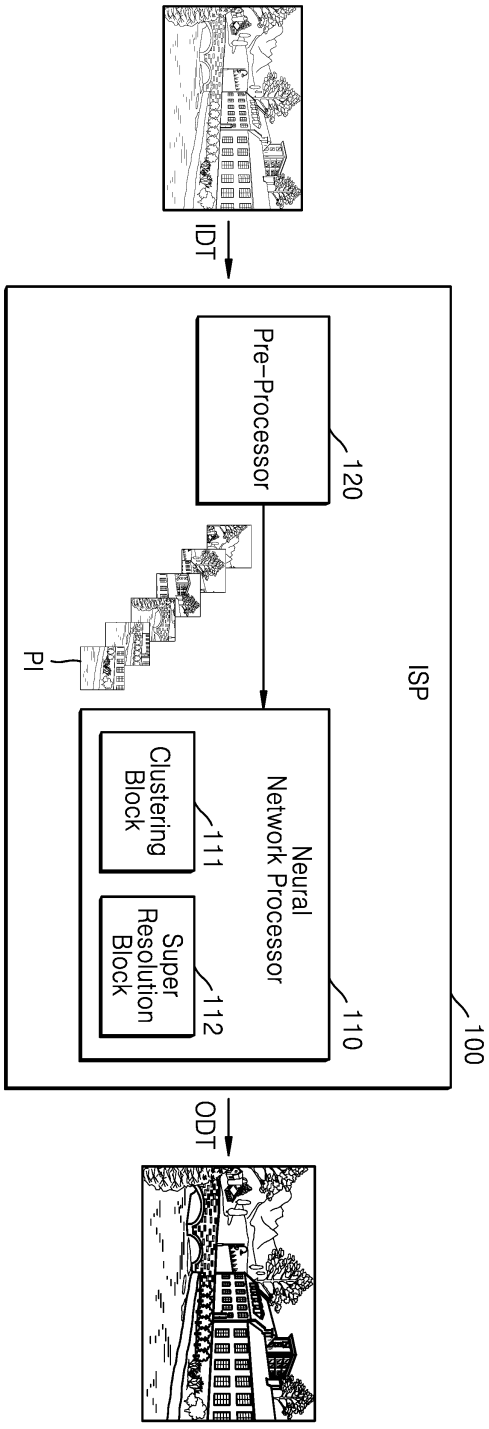
도면1


10
100
ISP
110
Neural Network Processor
200
CPU
300
RAM
700
400
Camera Module
500
Memory
600
Display

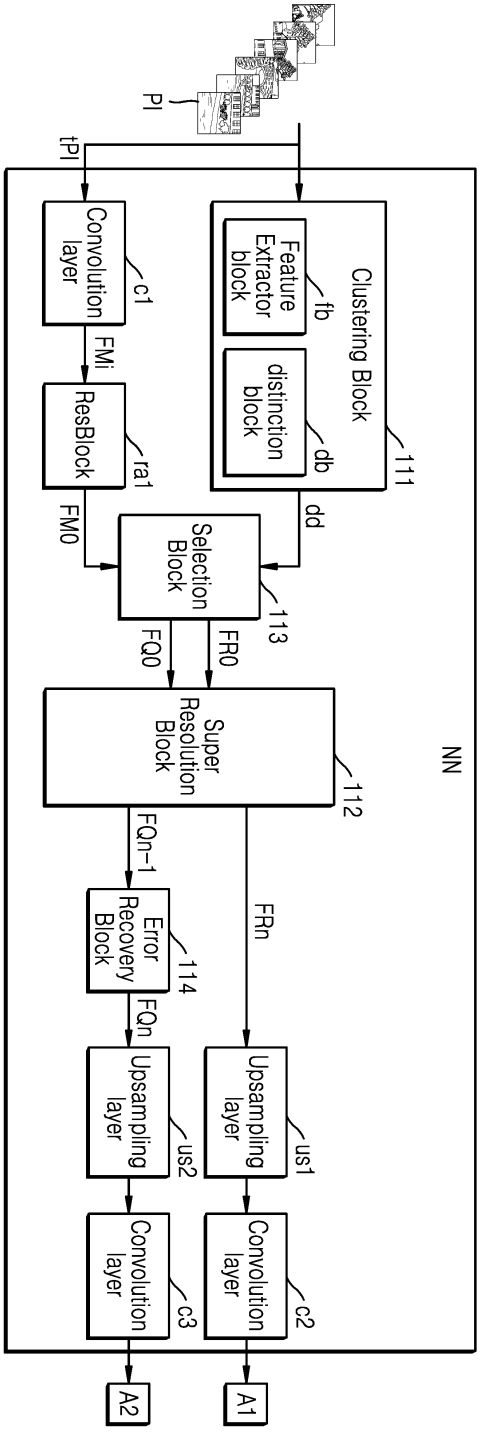
도면2



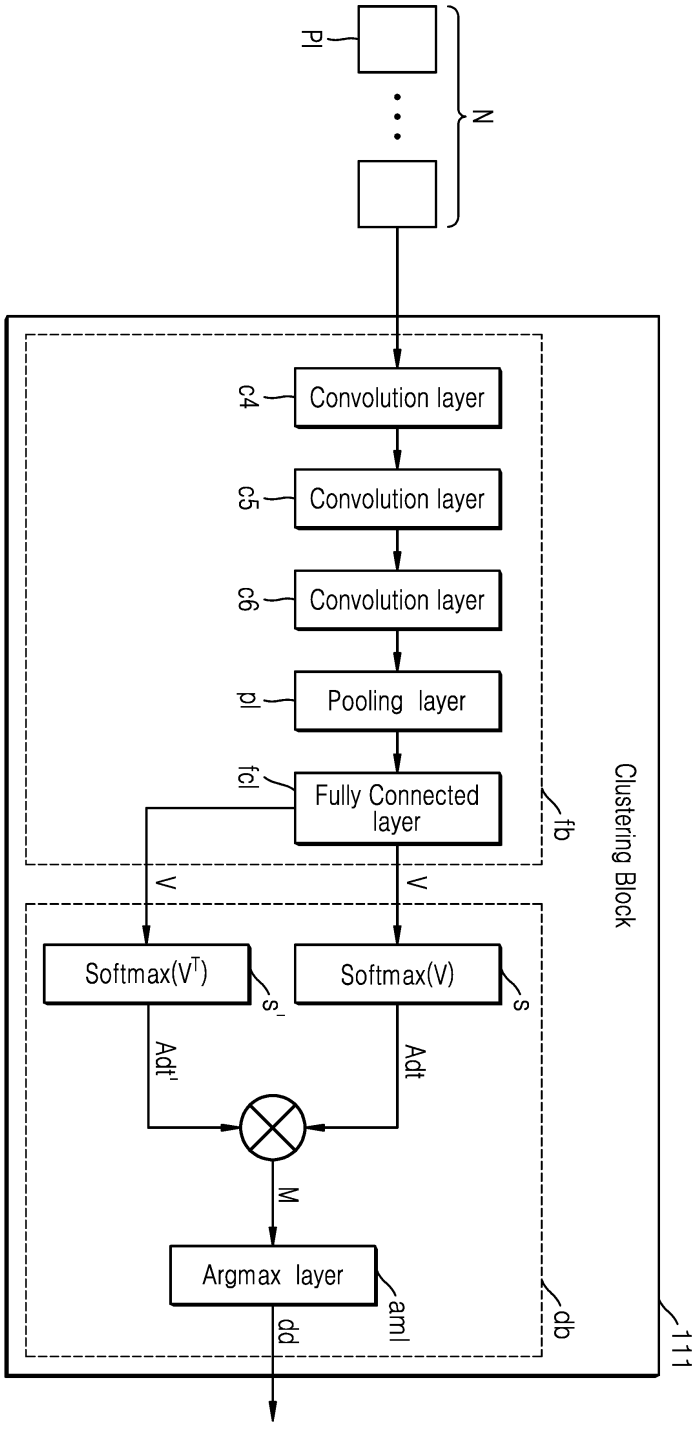
도면3


IDT
Pre-Processor
120
PI
ISP
Neural
Network Processor
Clustering
Block
111
Super
Resolution
Block
112
110
100
ODT

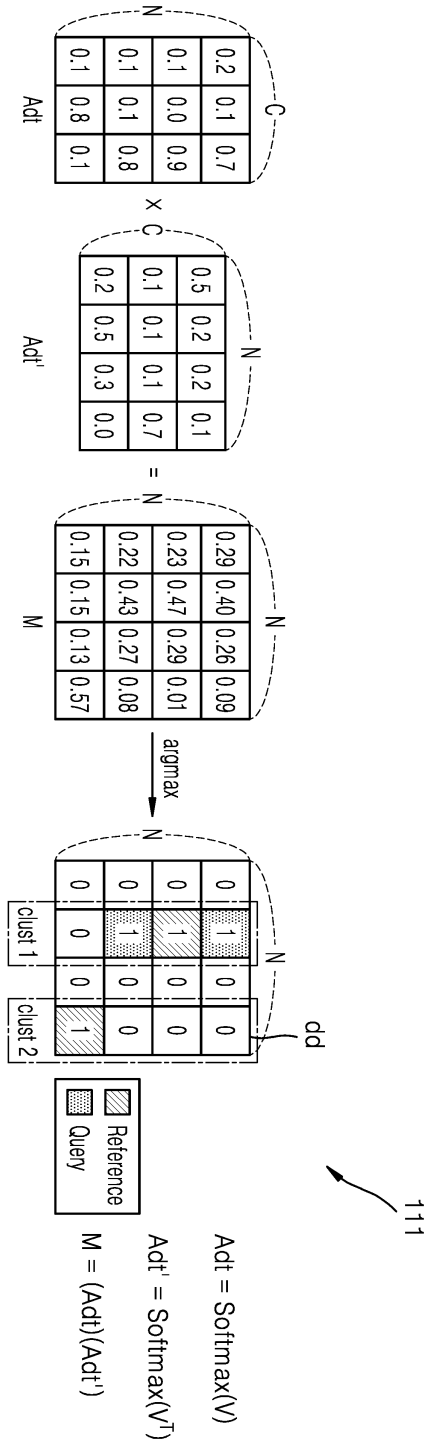
도면4


PI
tPI
Convolution layer
c1
FMi
ResBlock
ra1
FM0
Clustering Block
Feature Extractor block
fb
distinction block
db
111
dd
Selection Block
113
FQ0
FR0
Super Resolution Block
112
NN
FQn-1
Error Recovery Block
114
FQn
FRn
Upsampling layer
us2
Upsampling layer
us1
Convolution layer
c3
Convolution layer
c2
A2
A1

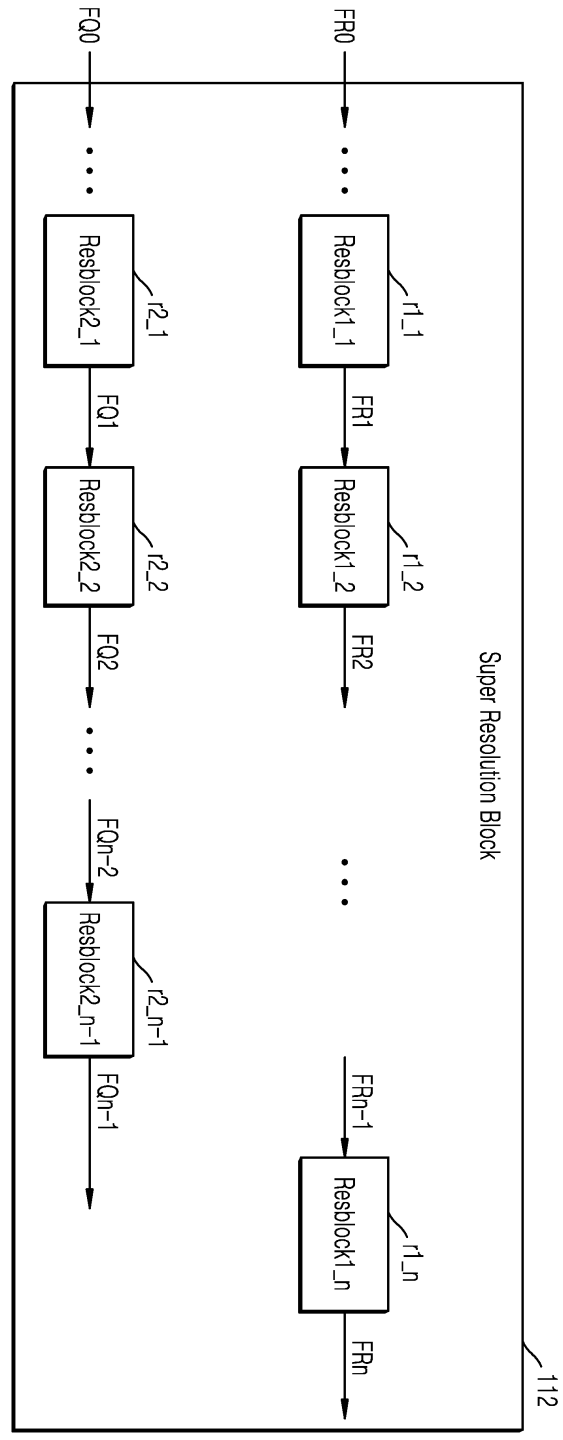
도면5


PI
N
c4
Convolution layer
c5
Convolution layer
c6
Convolution layer
pl
Pooling layer
fcl
Fully Connected layer
fb
Clustering Block
V
V
Softmax(V^T)
s'
Softmax(V)
s
Adt'
Adt
M
Argmax layer
aml
dd
db
111

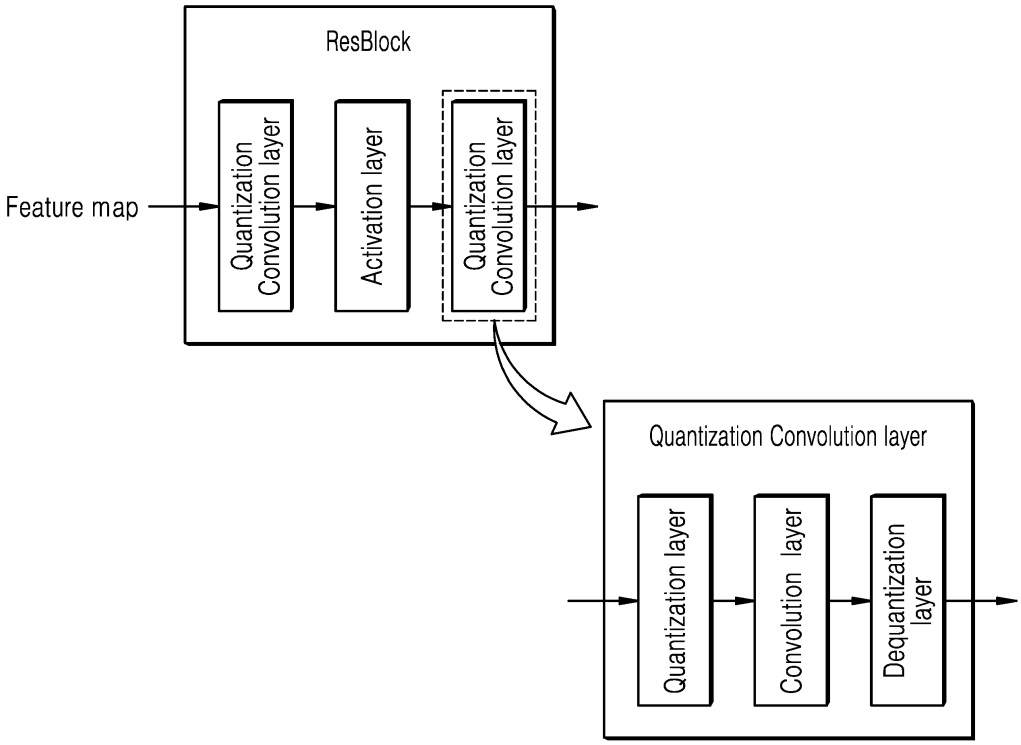
도면6


Adt
N
C
0.1 0.1 0.1 0.2
0.1 0.8 0.1 0.0 0.1
0.1 0.8 0.9 0.7
×
Adt'
C
N
0.2 0.1 0.5
0.2 0.5 0.1 0.1 0.2
0.5 0.3 0.1 0.1 0.2
0.0 0.7 0.1
=
M
N
N
0.15 0.22 0.23 0.29
0.15 0.43 0.47 0.40
0.13 0.27 0.29 0.26
0.57 0.08 0.01 0.09
argmax
clust 1
clust 2
N
N
0 0 0 0
0 1 1 1
0 0 0 0
1 0 0 0
dd
Query
Reference
111
Adt = Softmax(V)
Adt' = Softmax(V^T)
M = (Adt)(Adt')

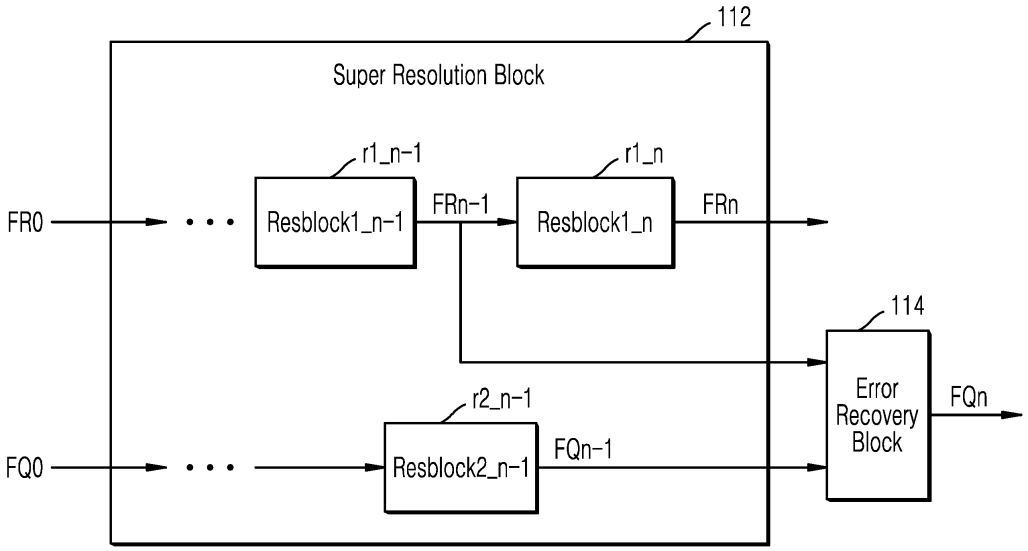
도면7


FQ0
FR0
Resblock2_1
r2_1
Resblock1_1
r1_1
FQ1
FR1
Resblock2_2
r2_2
Resblock1_2
r1_2
FQ2
FR2
Super Resolution Block
FQn-2
FRn-1
Resblock2_n-1
r2_n-1
Resblock1_n
r1_n
FQn-1
FRn
112

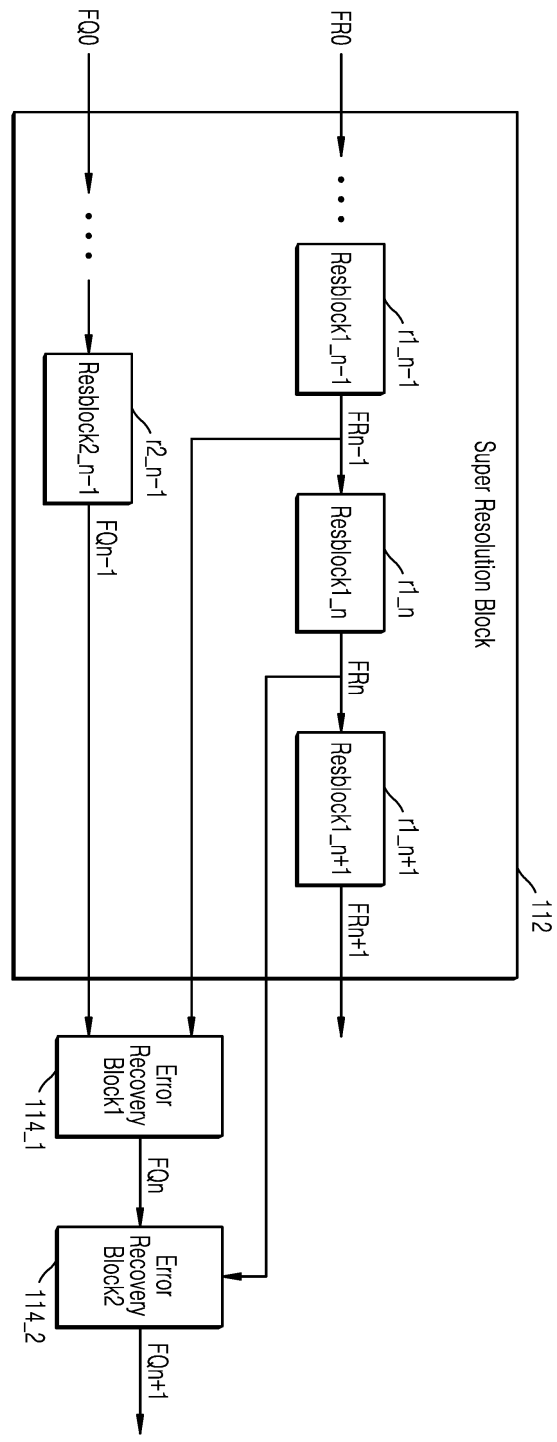
도면8


ResBlock
Feature map
Quantization Convolution layer
Activation layer
Quantization Convolution layer
Quantization Convolution layer
Quantization layer
Convolution layer
Dequantization layer

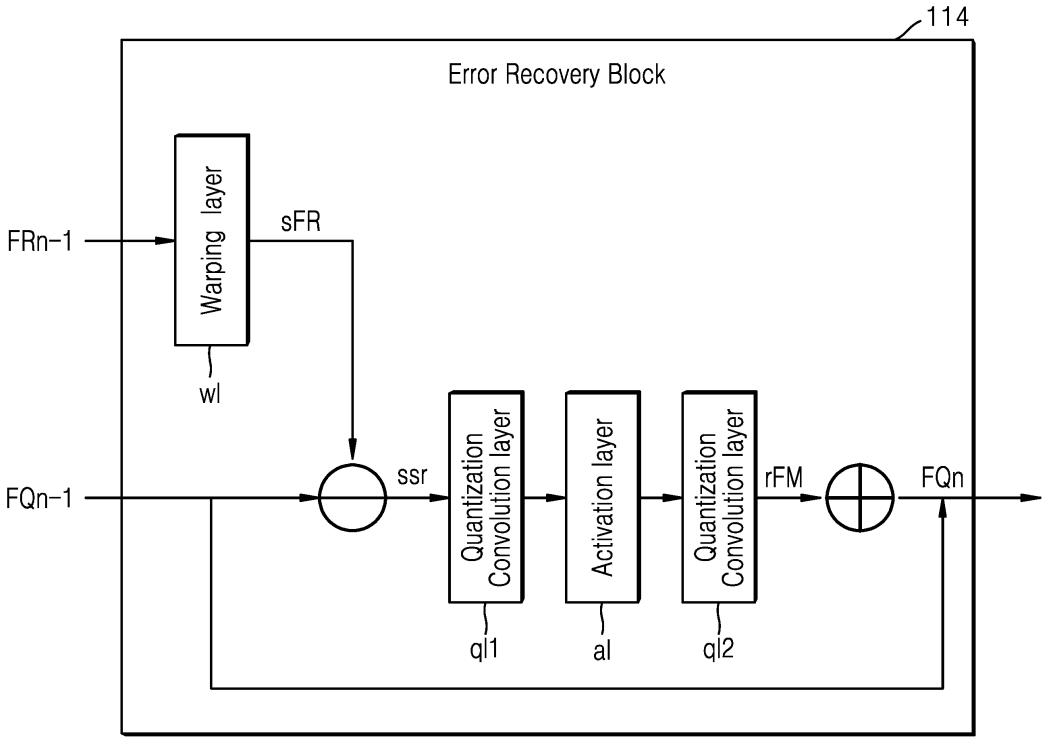
도면9a


112
Super Resolution Block
r1_n-1
r1_n
FR0
Resblock1_n-1
FRn-1
Resblock1_n
FRn
114
Error Recovery Block
FQn
r2_n-1
FQ0
Resblock2_n-1
FQn-1

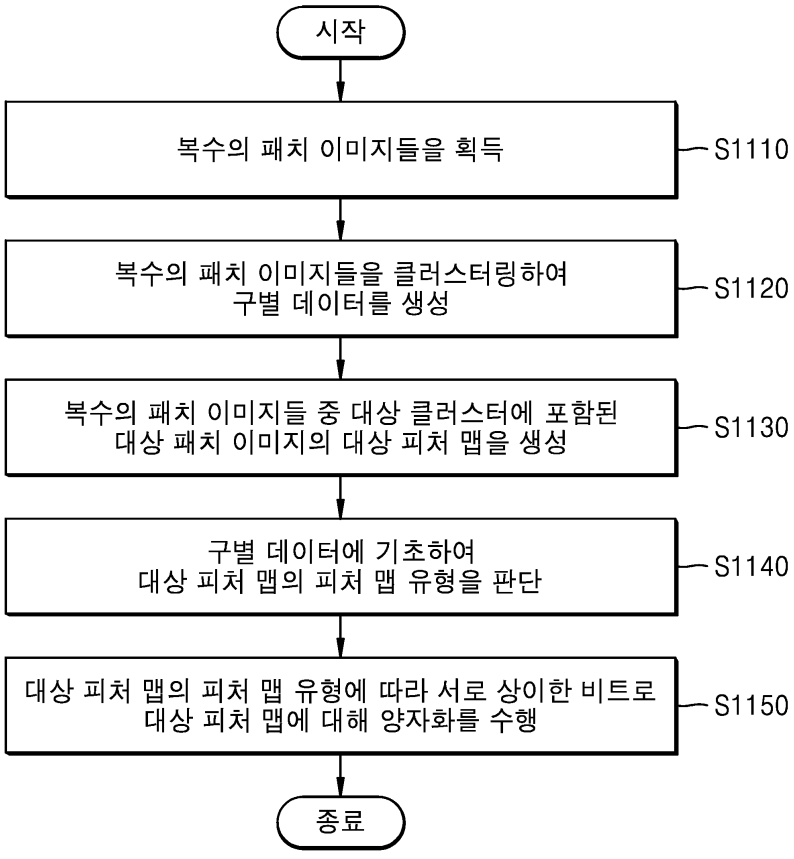
도면9b


FQ0
FR0
Resblock2_n-1
r2_n-1
FQn-1
Resblock1_n-1
r1_n-1
FRn-1
Resblock1_n
r1_n
FRn
Resblock1_n+1
r1_n+1
FRn+1
Super Resolution Block
112
Error Recovery Block1
114_1
FQn
Error Recovery Block2
114_2
FQn+1

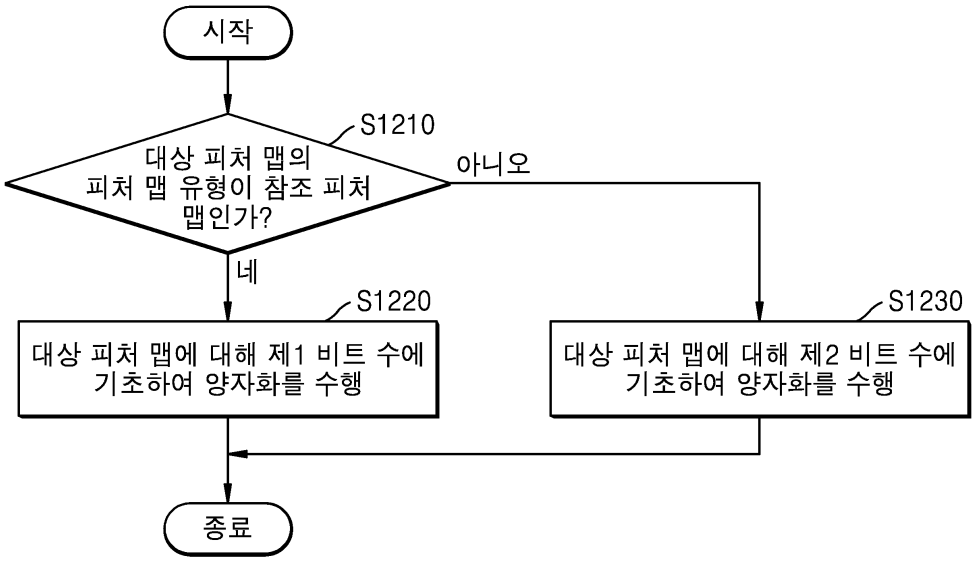
도면10


114
Error Recovery Block
FRn-1
Warping layer
wl
sFR
FQn-1
ssr
Quantization Convolution layer
ql1
Activation layer
al
Quantization Convolution layer
ql2
rFM
FQn

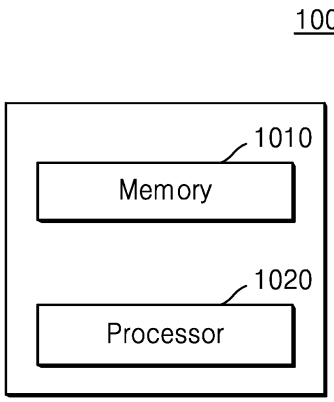
도면11


시작
복수의 패치 이미지들을 획득
S1110
복수의 패치 이미지들을 클러스터링하여 구별 데이터를 생성
S1120
복수의 패치 이미지들 중 대상 클러스터에 포함된 대상 패치 이미지의 대상 피처 맵을 생성
S1130
구별 데이터에 기초하여 대상 피처 맵의 피처 맵 유형을 판단
S1140
대상 피처 맵의 피처 맵 유형에 따라 서로 상이한 비트로 대상 피처 맵에 대해 양자화를 수행
S1150
종료

도면12


시작
S1210
대상 피처 맵의
피처 맵 유형이 참조 피처
맵인가?
아니오
네
S1220
대상 피처 맵에 대해 제1 비트 수에
기초하여 양자화를 수행
S1230
대상 피처 맵에 대해 제2 비트 수에
기초하여 양자화를 수행
종료

도면13


1000
1010
Memory
1020
Processor