



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0076800
(43) 공개일자 2020년06월30일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2006.01) G06F 17/15 (2006.01)
G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06F 17/153 (2013.01)
(21) 출원번호 10-2018-0165216
(22) 출원일자 2018년12월19일
심사청구일자 2018년12월19일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
박중선
서울특별시 서초구 신반포로 9 85동 403호 (반포
동, 주공아파트)
신동엽
서울특별시 강서구 초록마을로26길 49-11 301호
(화곡동, 두산빌라)
(74) 대리인
이기성

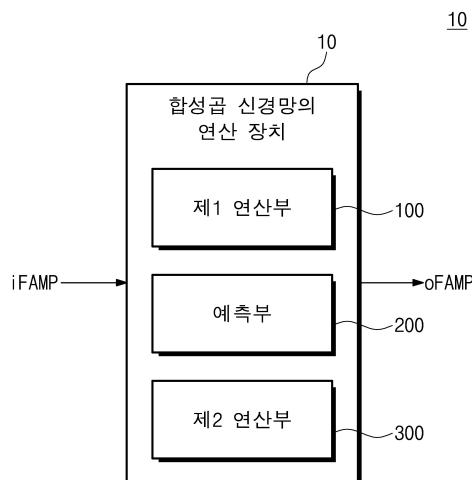
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 출력 특징 맵의 0에 대한 연산을 스킵할 수 있는 합성곱 신경망의 연산 장치 및 그 동작 방법

(57) 요약

본 출원의 일 실시예에 따르는 합성곱 신경망의 연산 장치는, 기설정된 모자이크 타입에 기초하여, 입력 특징 맵을 합성곱 연산하여 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 제1 연산부, 상기 적어도 하나의 픽셀정보에 기초하여, 상기 출력 특징 맵의 나머지 픽셀정보에 대해 스킵픽셀과 연산픽셀 중 어느 하나로 예측하는 예측부 및 상기 스킵픽셀에 대한 합성곱 연산을 스킵하여 0의 출력으로 처리하는 제2 연산부를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/04 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호 1711071760

부처명 과학기술정보통신부

연구관리전문기관 한국연구재단

연구사업명 개인기초연구(과기정통부)

연구과제명 사물인터넷용 저면적/저전력 전용 프로세서 설계를 위한 맞춤형 임베디드 메모리 개발

기 여 율 3.5/10

주관기관 고려대학교

연구기간 2018.04.01 ~ 2019.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711070699

부처명 과학기술정보통신부

연구관리전문기관 한국산업기술평가관리원

연구사업명 전자정보디바이스산업원천기술개발(정보화)

연구과제명 스마트 모바일 및 IoT 디바이스를 위한 뉴럴셀(Spiking Neural Cell)기반 SNP SoC 원천기

술 개발

기 여 율 4/10

주관기관 전자부품연구원

연구기간 2019.01.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711075738

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 정보통신기술인력양성(R&D)

연구과제명 스마트 농축산 IoA(Internet of Agriculture) 연구센터

기 여 율 2.5/10

주관기관 국립강원대학교 산학협력단

연구기간 2018.06.01 ~ 2019.02.28

명세서

청구범위

청구항 1

기설정된 모자이크 타입에 기초하여, 입력 특징 맵을 합성곱 연산하여 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 제1 연산부;

상기 적어도 하나의 픽셀정보에 기초하여, 상기 출력 특징 맵의 나머지 픽셀정보에 대해 스킵픽셀과 연산픽셀 중 어느 하나로 예측하는 예측부; 및

상기 스킵픽셀에 대한 합성곱 연산을 스킵하여 0의 출력으로 처리하는 제2 연산부를 포함하는 합성곱 신경망의 연산 장치.

청구항 2

제1항에 있어서,

상기 제1 연산부는, 기설정된 모자이크 타입에 기초하여, 미리 정해진 예측 패턴을 검출하는 합성곱 신경망의 연산 장치.

청구항 3

제2항에 있어서,

상기 제1 연산부는, 상기 예측 패턴에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵에 대한 제1 연산 순서정보를 결정하고,

상기 제1 연산순서정보는, 상기 예측 패턴에 따라 결정되는 상기 스트라이드에 대한 동작순서정보인 합성곱 신경망의 연산 장치.

청구항 4

제3항에 있어서,

상기 제1 연산부는, 제1 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하여 상기 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 합성곱 신경망의 연산 장치.

청구항 5

제3항에 있어서,

상기 제1 연산부는, 상기 입력 특징 맵의 최상위비트와 상기 커널에 기설정된 필터계수를 합성곱 연산할 때,

상기 입력 특징 맵의 하위비트들과 상기 기설정된 필터계수를 합성곱 연산하고, 연산된 연산결과를 보정상수로 가중하는 합성곱 신경망의 연산 장치.

청구항 6

제1항에 있어서,

상기 출력 특징 맵의 0의 출력에 기초하여, 새로 생성되는 다음의 입력 특징 맵을 합성곱 연산하는 희소 행렬 가속부를 더 포함하고,

상기 희소 행렬 가속부는, 상기 출력 특징 맵 중 상기 0의 출력을 제외한 픽셀정보에 대응하는 커널과 스트라이드를 통해 상기 다음의 입력 특징 맵을 합성곱 연산하는 합성곱 신경망의 연산 장치.

청구항 7

제1항에 있어서,

상기 예측부는, 상기 적어도 하나의 픽셀정보의 0의 출력 패턴과 상기 예측 패턴 간의 동일 여부에 따라, 상기 예측 패턴을 따라 기설정된 상기 나머지 출력정보의 제1 영역을 상기 스킵픽셀로 예측하는 합성곱 신경망의 연산 장치.

청구항 8

제2항에 있어서,

상기 적어도 하나의 픽셀정보는, 상기 예측 패턴을 따라 연산된 C_{NM} 의 픽셀 행렬이고,

여기서, C는 상기 출력 특징 맵에 대응되는 픽셀 행렬이고, N은 상기 픽셀 행렬의 X축 위치정보이고, M은 상기 픽셀 행렬의 Y축 위치정보인 합성곱 신경망의 연산 장치.

청구항 9

합성곱 신경망의 연산 장치의 동작방법으로서,

기설정된 모자이크 타입에 기초하여, 입력 특징 맵을 합성곱 연산하고, 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 단계;

상기 적어도 하나의 픽셀정보에 기초하여, 상기 출력 특징 맵의 나머지 픽셀정보에 대해 스킵픽셀과 연산픽셀 중 어느 하나로 예측하는 단계; 및

상기 스킵픽셀에 대한 합성곱 연산을 스킵하여 0의 출력으로 처리하는 단계를 포함하는 합성곱 신경망의 연산 장치의 동작방법.

청구항 10

제9항에 있어서,

상기 연산픽셀에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵을 합성곱 연산하여 연산픽셀에 대한 컨볼루션 정보를 출력하는 단계; 및

상기 0의 출력, 상기 적어도 하나의 픽셀정보 및 상기 연산픽셀에 대한 컨볼루션 정보를 병합하여, 상기 출력 특징 맵을 출력하는 단계를 더 포함하는 합성곱 신경망의 연산 장치의 동작방법.

청구항 11

제9항에 있어서,

상기 적어도 하나의 픽셀정보를 출력하는 단계는, 상기 기설정된 모자이크 타입에 기초하여, 미리 정해진 예측 패턴을 검출하는 단계;

상기 예측 패턴에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵에 대한 제1 연산순서정보를 결정하는 단계; 및

상기 제1 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하는 단계를 포함하는 합성곱 신경망의 연산 장치의 동작방법.

청구항 12

제10항에 있어서,

상기 예측하는 단계는, 상기 적어도 하나의 픽셀정보와 상기 예측 패턴을 비교하는 단계;

상기 적어도 하나의 픽셀정보와 상기 예측 패턴 간의 동일 여부에 따라, 상기 나머지 픽셀정보 중 상기 예측 패턴에 따라 기설정된 제1 영역을 상기 스킵픽셀로 예측하는 단계; 및

상기 나머지 픽셀정보 중 나머지 영역을 상기 연산픽셀로 예측하는 단계를 포함하는 합성곱 신경망의 연산 장치의 동작방법.

청구항 13

제11항에 있어서,

상기 0의 출력으로 처리하는 단계는, 상기 연산픽셀에 대응하는 커널과 스트라이드를 통해 상기 입력 특징 맵에 대한 제2 연산순서정보를 결정하는 단계;

상기 제2 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하여 상기 연산픽셀에 대한 컨볼루션 정보를 출력하는 단계;

상기 연산픽셀에 대한 컨볼루션 정보에 기초하여, 상기 나머지 픽셀정보를 출력하는 단계; 및

상기 적어도 하나의 픽셀정보와 상기 나머지 픽셀정보를 병합하여 상기 출력 특징 맵을 출력하는 단계를 포함하는 합성곱 신경망의 연산 장치의 동작방법.

발명의 설명

기술 분야

[0001] 본 출원은 합성곱 신경망의 연산 장치 및 그 동작 방법에 관한 것으로, 출력 특징 맵에서 0의 출력을 예측하여, 입력 특징 맵의 일부 연산을 스킵하는 합성곱 신경망의 연산 장치 및 그 동작 방법에 관한 것이다.

배경 기술

[0002] 최근 인공 신경망에 관한 기술이 각광을 받고 있다. 딥러닝(deep learning) 기술의 핵심 계산 모형인 컨볼루션 뉴럴 네트워크(Convolutional Neural Network, 이하에서 "CNN"이라고 지칭함)은, 각각의 뉴런이 인간의 시각 신경계에서 중복 영역의 응답 특성과 유사한 특성을 갖도록 다중 계층으로 배치된 인공신경망의 한 종류이다.

[0003] 이러한 CNN은 계층의 수가 많아질수록 복잡한 패턴의 영상 및 음성을 높은 인식률로 인식할 수 있다는 장점이 있다. 그에 따라, 오랫동안 정체 상태에 머물러 있던 영상 및 음성 인식률이 최근 들어 CNN 기술에 의해 극적으로 높아지고, ImageNet과 같은 각종 영상인식 벤치마킹시험에서 CNN 시스템들이 최고기록을 경신하고 있다. 그 결과, CNN은 영상인식, 음성인식, 언어번역 등 다양한 머신러닝(machine learning) 분야에서 각광받고 있다.

[0004] 특히, CNN은 여러 단계의 컨볼루션 레이어들을 계산하기 때문에, 전력 에너지 소모가 상당하다. 또한, 컨볼루션 레이어는 입력 특징 맵과 필터 계수 사이의 합성곱(Convolution) 연산을 통해 출력되는 컨볼루션 연산결과와 컨볼루션 연산결과를 비선형 활성화 함수(Non linear activation function)를 통해 활성화하여 출력되는 출력 특징 맵을 포함하고 있다.

[0005] 이에 따라, 본 출원에서는, CNN을 이용하는 하드웨어 장치에서, 컨볼루션 레이어들을 계산하는 데 소모되는 에너지 소모를 줄일 수 있는 방안을 제안하고 자 한다.

발명의 내용

해결하려는 과제

[0006] 본 출원의 목적은, 출력 특징 맵에서 0의 출력을 정확하게 예측하고, 0의 출력에 대응되는 입력 특징 맵의 픽셀들에 대한 연산을 스킵할 수 있는 합성곱 신경망의 연산 장치 및 그 동작 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0007] 본 출원의 일 실시예에 따르는 합성곱 신경망의 연산 장치는, 기설정된 모자이크 타입에 기초하여, 입력 특징 맵을 합성곱 연산하여 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 제1 연산부, 상기 적어도 하나의 픽셀정보에 기초하여, 상기 출력 특징 맵의 나머지 픽셀정보에 대해 스킵픽셀과 연산픽셀 중 어느 하나로 예측하는 예측부 및 상기 스킵픽셀에 대한 합성곱 연산을 스킵하여 0의 출력으로 처리하는 제2 연산부를 포함한다.

[0008] 실시예에 있어서, 상기 제1 연산부는, 기설정된 모자이크 타입에 기초하여, 미리 정해진 예측 패턴을 검출한다.

[0009] 실시예에 있어서, 상기 제1 연산부는, 상기 예측 패턴에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵에 대한 제1 연산순서정보를 결정하고, 상기 제1 연산순서정보는, 상기 예측 패턴에 따라 결정되는 상기 스트

라이드에 대한 동작순서정보이다.

- [0010] 실시예에 있어서, 상기 제1 연산부는, 제1 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하여 상기 출력 특징 맵의 적어도 하나의 픽셀정보를 출력한다.
- [0011] 실시예에 있어서, 상기 제1 연산부는, 상기 입력 특징 맵의 최상위비트와 상기 커널에 기설정된 필터계수를 합성곱 연산할 때, 상기 입력 특징 맵의 하위비트들과 상기 기설정된 필터계수를 합성곱하여 연산된 연산결과를 보정상수로 가중한다.
- [0012] 실시예에 있어서, 상기 출력 특징 맵의 0의 출력에 기초하여, 다음의 입력 특징 맵을 합성곱 연산하는 최소 행렬 가속부를 더 포함하고, 상기 최소 행렬 가속부는, 상기 0의 출력이 아닌 픽셀정보에 대응하는 커널과 스트라이드를 통해 다음의 입력 특징 맵을 연산한다.
- [0013] 실시예에 있어서, 상기 예측부는, 상기 적어도 하나의 픽셀정보의 0의 출력 패턴과 상기 예측 패턴 간의 동일 여부에 따라, 상기 예측 패턴을 따라 기설정된 상기 나머지 출력정보의 제1 영역을 상기 스킵픽셀로 예측한다.
- [0014] 실시예에 있어서, 상기 적어도 하나의 픽셀정보는, 상기 예측 패턴을 따라 연산된 C_{NM} 의 픽셀 행렬이고, 여기서, C는 상기 출력 특징 맵에 대응되는 픽셀 행렬이고, N은 상기 픽셀 행렬의 X축 위치정보이고, M은 상기 픽셀 행렬의 Y축 위치정보이다.
- [0015] 실시예에 있어서, 상기 예측 패턴은 C11, C13, C15, C31, C33, C35, C51, C53, 및 C55 위치에서 픽셀정보를 가지는 행렬이다.
- [0016] 실시예에 있어서, 상기 예측 패턴은 C11, C14, C22, C25, C33, C36, C41, C44, C52, C55, C63 및 C66 위치에서 픽셀정보를 가지는 행렬이다.
- [0017] 실시예에 있어서, 상기 예측 패턴은 C11, C13, C15, C22, C24, C26, C31, C33, C35, C42, C44, C46, C51, C53, C55, C62, C64 및 C66 위치에서 픽셀정보를 가지는 행렬이다.
- [0018] 실시예에 있어서, 상기 예측 패턴은 C11, C14, C22, C25, C33, C36, C43, C46, C52, C55, C61 및 C64 위치에서 픽셀정보를 가지는 행렬이다.
- [0019] 실시예에 있어서, 상기 예측 패턴은 C11, C13, C15, C22, C24, C26, C32, C34, C36, C41, C43, C45, C51, C53, C55, C62, C64 및 C66 위치에서 픽셀정보를 가지는 행렬이다.
- [0020] 실시예에 있어서, 상기 예측 패턴은 C11, C12, C14, C15, C22, C23, C25, C26, C31, C33, C34, C36, C41, C43, C44, C46, C52, C53, C55, C56, C61, C62, C64 및 C65 위치에서 픽셀정보를 가지는 행렬이다.
- [0021] 실시예에 있어서, 상기 예측 패턴은 C11, C12, C14, C15, C22, C23, C25, C26, C31, C33, C34, C36, C41, C42, C44, C45, C52, C53, C55, C56, C61, C62, C64 및 C65 위치에서 픽셀정보를 가지는 행렬이다.
- [0022] 본 출원의 실시예에 따르는 합성곱 신경망의 연산 장치의 동작방법으로서, 기설정된 모자이크 타입에 기초하여, 입력 특징 맵을 합성곱 연산하고, 출력 특징 맵의 적어도 하나의 픽셀정보를 출력하는 단계, 상기 적어도 하나의 픽셀정보에 기초하여, 상기 출력 특징 맵의 나머지 픽셀정보에 대해 스킵픽셀과 연산픽셀 중 어느 하나로 예측하는 단계 및 상기 스킵픽셀에 대한 합성곱 연산을 스킵하여 0의 출력으로 처리하는 단계를 포함한다.
- [0023] 실시예에 있어서, 상기 연산픽셀에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵을 합성곱 연산하여 연산픽셀에 대한 컨볼루션 정보를 출력하는 단계 및 상기 0의 출력, 상기 적어도 하나의 픽셀정보 및 상기 연산픽셀에 대한 컨볼루션 정보를 병합하여, 상기 출력 특징 맵을 출력하는 단계를 더 포함한다.
- [0024] 실시예에 있어서, 상기 적어도 하나의 픽셀정보를 출력하는 단계는, 상기 기설정된 모자이크 타입에 기초하여, 미리 정해진 예측 패턴을 검출하는 단계, 상기 예측 패턴에 대응하는 커널과 스트라이드를 통해, 상기 입력 특징 맵에 대한 제1 연산순서정보를 결정하는 단계 및 상기 제1 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하는 단계를 포함한다.
- [0025] 실시예에 있어서, 상기 예측하는 단계는, 상기 적어도 하나의 픽셀정보와 상기 예측 패턴을 비교하는 단계, 상기 적어도 하나의 픽셀정보와 상기 예측 패턴 간의 동일 여부에 따라, 상기 나머지 픽셀정보 중 상기 예측 패턴에 따라 기설정된 제1 영역을 상기 스킵픽셀로 예측하는 단계 및 상기 나머지 픽셀정보 중 나머지 영역을 상기 연산픽셀로 예측하는 단계를 포함한다.

[0026] 실시예에 있어서, 상기 0의 출력으로 처리하는 단계는, 상기 연산픽셀에 대응하는 커널과 스트라이드를 통해 상기 입력 특징 맵에 대한 제2 연산순서정보를 결정하는 단계, 상기 제2 연산순서정보에 따라, 상기 입력 특징 맵과 상기 커널을 합성곱 연산하여 상기 연산픽셀에 대한 컨볼루션 정보를 출력하는 단계, 상기 연산픽셀에 대한 컨볼루션 정보에 기초하여, 상기 나머지 픽셀정보를 출력하는 단계 및 상기 적어도 하나의 픽셀정보와 상기 나머지 픽셀정보를 병합하여 상기 출력 특징 맵을 출력하는 단계를 포함한다.

발명의 효과

[0027] 본 출원의 실시 예에 따른 합성곱 신경망의 연산 장치 및 그 동작 방법은 입력 특징 맵에 대한 합성곱 연산에 따른 시스템부하 및 전력소모를 감소시킬 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 출원의 일 실시 예에 따르는 합성곱 신경망의 연산 장치의 블록도이다.

도 2는 도 1의 합성곱 신경망의 연산 장치의 동작의 일 예를 보여주는 도면이다.

도 3은 도 1의 연산부의 동작에 따른 실시 예이다.

도 4는 도 1의 제1 연산부에 대한 실시 예이다.

도 5a 내지 도 5G는 기설정된 모자이크 타입에 따른 예측 패턴의 실시 예들이다.

도 6은 도 1의 제2 연산부에 대한 실시 예이다.

도 7은 본 출원의 다른 실시예에 따르는 합성곱 신경망의 연산 장치의 블록도이다.

도 8은 합성곱 신경망의 컨볼루션 시스템에 대한 실시 예이다.

도 9는 본 출원의 실시예에 따른 합성곱 신경망의 연산 장치의 동작 프로세스이다.

도 10은 도 1의 제1 연산부에 대한 동작 프로세스이다.

도 11은 도 1의 예측부에 대한 동작 프로세스이다.

도 12는 도 1의 제2 연산부에 대한 동작 프로세스이다.

발명을 실시하기 위한 구체적인 내용

[0029] 본 명세서에 개시되어 있는 본 출원의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 출원의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 출원의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0030] 본 출원의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시 예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 출원의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 출원의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0031] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 출원의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1구성요소는 제2구성요소로 명명될 수 있고, 유사하게 제2구성요소는 제1구성요소로도 명명될 수 있다.

[0032] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0033] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 출원을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서,

"포함하다" 또는 "가지다" 등의 용어는 실시된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

- [0034] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 출원이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0035] 이하, 첨부한 도면을 참조하여 본 출원의 바람직한 실시 예를 설명함으로써, 본 출원을 상세히 설명될 것이다.
- [0036] 도 1은 본 출원의 일 실시 예에 따르는 합성곱 신경망의 연산 장치(10)의 블록도이고, 도 2는 도 1의 합성곱 신경망의 연산 장치(10)의 동작의 일 예를 보여주는 도면이다.
- [0037] 도 1과 도 2를 참조하면, 합성곱 신경망의 연산 장치(10)는 제1 연산부(100), 예측부(200) 및 제2 연산부(300)를 포함할 수 있다.
- [0038] 먼저, 제1 연산부(100)는 기설정된 모자이크 타입(Mosaic Type, 106)에 기초하여, 입력 특징 맵(input feature map, "iFMAP", 이하, 101)을 합성곱 연산할 수 있다. 여기서, 기설정된 모자이크 타입(106)은 출력 특징 맵(301)의 인접한 0의 픽셀을 예측하기 위한 예측 패턴(예컨대, 106_1~106_7)을 정의한 픽셀 행렬일 수 있다.
- [0039] 도 2에 도시된 바와 같이, 이때, 입력 특징 맵(101)은 픽셀 행렬의 형태를 갖는 입력데이터일 수 있다. 예를 들면, 제어장치나 중앙처리장치는 입력 특징 맵(101)을 생성 또는 전송받아, 메모리, 저장장치나 클라우드 서버에 입력 특징 맵(101)을 저장할 수 있다. 즉, 제1 연산부(100)는 메모리나 클라우드 서버에 저장된 또는 제어장치나 중앙처리장치에서 생성된 입력 특징 맵(101)을 입력받아 합성곱 연산할 수 있다.
- [0040] 실시예에 따른 제1 연산부(100)는 기설정된 모자이크 타입(106)에 기초하여, 입력 특징 맵(101)을 합성곱 연산하고, 출력 특징 맵(output feature map, oFMAP, 301)의 적어도 하나의 픽셀(102_1~102_N)을 출력할 수 있다. 여기서, 기설정된 모자이크 타입(106)의 행렬크기는 출력 특징 맵(301)에 대응되는 크기일 수 있다. 이하, 도 5a 내지 도 5g에서, 기설정된 모자이크 타입(106)에 따른 예측 패턴(예컨대, 106_1~106_7)에 대해 보다 구체적으로 설명될 것이다.
- [0041] 보다 구체적으로, 제1 연산부(100)는 기설정된 모자이크 타입(106)에 기초하여, 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)을 검출할 수 있다. 예를 들면, 도 2에 도시된 바와 같이, 제1 연산부(100)는 기설정된 모자이크 타입(106)에 기초하여, 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_3)을 격자 패턴으로 검출할 수 있다.
- [0042] 그런 다음, 제1 연산부(100)는 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)에 대응하는 커널(Kernel, 103)과 스트라이드(Stride)를 통해, 입력 특징 맵(101)에 대한 제1 연산순서정보를 결정할 수 있다.
- [0043] 여기서, 커널(121)은 입력 특징 맵(101)을 필터링 합성곱 연산하기 위한 픽셀 행렬이다. 또한, 스트라이드는 출력 특징 맵(301)의 픽셀 수를 변경시키기 위한 동작 정보이다. 그리고, 제1 연산순서정보는 예측 패턴(예컨대, 106_1~106_7)에 따라 커널(121)과 입력 특징 맵(101)을 순서대로 합성곱 연산하기 위하여, 스트라이드에 대한 동작순서정보일 수 있다. 즉, 제1 연산순서정보는 출력 특징 맵(301)의 하나의 픽셀정보에 대하여, 입력 특징 맵(101)과 커널(121) 간의 연산순서를 대응시킨 정보일 수 있다.
- [0044] 이하, 도 3을 참조하여, 제1 연산순서정보에 대해 보다 구체적으로 설명될 것이다.
- [0045] 도 3은 도 1의 연산부(100)의 동작의 일 예를 보여주는 도면이다.
- [0046] 도 3에 도시된 바와 같이, 입력 특징 맵(101)이 7 X 7 크기의 픽셀 행렬이고, 커널(121)이 3 X 3의 크기의 픽셀 행렬이며, 스트라이드가 1이고, 출력 특징 맵(301)이 5 X 5 크기의 픽셀 행렬이며, 기설정된 모자이크 타입(106)에 따라, 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_3)이 격자 패턴인 경우, 제1 연산부(100)는 제1 연산순서정보를 결정할 수 있다.
- [0047] 보다 구체적으로, 제1 연산부(100)는 격자 패턴에 해당하는 스트라이드 순서마다 입력 특징 맵(101)과 커널(121)을 합성곱 연산하는 제1 연산순서정보를 결정할 수 있다. 예를 들면, 제1 연산순서정보는 격자 패턴에 해당하는 1크기의 스트라이드 1번째, 1크기의 스트라이드 3번째, 1크기의 스트라이드 5번째 등과 같은 순서정보를

포함할 수 있다. 즉, 제1 연산순서정보는 격자 패턴에 해당하는 13개의 스트라이드의 순서정보를 포함할 수 있다.

- [0048] 다시 도 1과 도 2를 참조하면, 제1 연산부(100)는 기설정된 모자이크 타입(106)의 예측 패턴(예컨대, 106_1~106_7)에 대응하는 커널(121)과 스트라이드를 통해, 입력 특징 맵(101)에 대한 제1 연산순서정보를 가변시킬 수 있다. 즉, 입력 특징 맵(101)에 대한 제1 연산순서정보는 예측 패턴(예컨대, 106_1~106_7)에 따라, 스트라이드의 동작순서가 가변될 수 있다.
- [0049] 이때, 제1 연산부(100)는 제1 연산순서정보에 따라, 입력 특징 맵(101)을 합성곱 연산하여, 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다.
- [0050] 실시예에 따라, 제1 연산부(100)는 입력 특징 맵(101)을 합성곱 연산할 때, 입력 특징 맵(101)의 하위비트들과 기설정된 필터계수를 합성곱 연산하여 계산된 연산결과를 보정상수로 결정할 수 있다. 이때, 제1 연산부(100)는 입력 특징 맵(101)의 최상위비트와 커널(121)에 기설정된 필터계수를 보정상수로 가중할 수 있다.
- [0051] 즉, 제1 연산부(100)는 제1 연산순서정보에 따라, 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)에 대응하는 입력 특징 맵(101)과 커널(121)을 합성곱 연산하고, 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다.
- [0052] 여기서, 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)는 기설정된 모자이크 타입(106)의 예측 패턴(예컨대, 106_1~106_7)을 따라, 합성곱 연산된 C_{NM} 의 픽셀 행렬로서, C는 출력 특징 맵(301)에 대한 픽셀 행렬을 의미하고, N은 상기 C_{NM} 의 픽셀 행렬의 X축 위치정보이고, M은 상기 C_{NM} 의 픽셀 행렬의 Y축 위치정보일 수 있다.
- [0053] 다음으로, 예측부(200)는 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)에 기초하여, 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)에 대해 스킵픽셀(104_1~104_N)과 연산픽셀(105_1~105_N) 중 어느 하나로 예측할 수 있다.
- [0054] 보다 구체적으로, 예측부(200)는 제1 연산부(100)로부터 적어도 하나의 픽셀정보(102_1~102_N)와 기설정된 모자이크 타입(106)에 따른 예측 패턴(예컨대, 106_1~106_7)을 전송받을 수 있다.
- [0055] 그런 다음, 예측부(200)는 적어도 하나의 픽셀정보(102_1~102_N)와 예측 패턴(예컨대, 106_1~106_7)을 비교할 수 있다. 예를 들면, 예측부(200)는 적어도 하나의 픽셀정보(102_1~102_N) 중 중 0의 픽셀정보와 예측 패턴(예컨대, 106_1~106_7)의 0의 픽셀정보 간의 동일 여부를 판단할 수 있다.
- [0056] 이때, 적어도 하나의 픽셀정보(102_1~102_N)와 예측 패턴(예컨대, 106_1~106_7) 간의 동일 여부에 따라, 예측부(200)는 나머지 픽셀정보(103_1~103_N) 중 예측 패턴(예컨대, 106_1~106_7)에 따라 기설정된 제1 영역(104)에 위치하는 픽셀정보를 스킵픽셀(104_1~104_N)로 예측할 수 있다. 즉, 적어도 하나의 픽셀정보(102_1~102_N) 중 중 0의 픽셀정보와 예측 패턴(예컨대, 106_1~106_7)의 0의 픽셀정보가 동일한 경우, 예측부(200)는 나머지 픽셀정보(103_1~103_N) 중 예측 패턴(예컨대, 106_1~106_7)에 따라 기설정된 제1 영역(104)에 위치하는 픽셀정보를 스킵픽셀(104_1~104_N)로 예측할 수 있다.
- [0057] 또한, 예측부(200)는 나머지 픽셀정보(103_1~103_N) 중 나머지 영역(105)에 위치하는 픽셀정보를 연산픽셀(105_1~105_N)로 예측할 수 있다. 여기서, 제1 영역(104)과 나머지 영역(105)는 서로 인접한 영역일 수 있다. 즉, 예측부(200)는 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)에 기초하여, 나머지 픽셀정보(103_1~103_N)에 대해 스킵픽셀(104_1~104_N)과 연산픽셀(105_1~105_N) 중 어느 하나로 예측할 수 있다.
- [0058] 예를 들면, 도 2에 도시된 바와 같이, 적어도 하나의 픽셀정보(102_1~102_N) 중 0의 픽셀정보와 예측패턴(예컨대, 105_3)이 동일한 경우, 예측부(200)는 나머지 픽셀정보(103_1~103_N) 중 예측패턴(예컨대, 105_3)에 따라 기설정된 제1 영역(104)에 위치하는 픽셀정보를 스킵픽셀(104_1~104_N)로 예측할 수 있다. 이때, 예측부(200)는 나머지 영역(105)에 위치하는 픽셀정보를 연산픽셀(105_1~105_N)로 예측할 수 있다.
- [0059] 또한, 적어도 하나의 픽셀정보(102_1~102_N) 중 0의 픽셀정보와 예측패턴(예컨대, 105_3)이 서로 동일하지 않은 경우, 예측부(200)는 나머지 픽셀정보(103_1~103_N) 중 제1 영역(104)과 나머지 영역(105)에 위치하는 픽셀정보를 연산픽셀(105_1~105_N)로 예측할 수 있다.
- [0060] 실시예에 따른 제2 연산부(300)는 예측부(200)를 통해 예측된 스킵픽셀(104_1~104_N)에 대한 합성곱 연산을 스킵하여 0의 출력(예컨대, 302)으로 처리할 수 있다. 보다 구체적으로, 제2 연산부(300)는 예측부(200)를 통해

나머지 픽셀정보(103_1~103_N) 중 어느 하나가 스킵픽셀(104_1~104_N)로 예측된 경우, 어느 하나에 대한 연산을 스킵하고, 0의 출력(예컨대, 302)으로 처리할 수 있다. 즉, 제2 연산부(300)는 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N) 중 스킵픽셀(104_1~104_N)을 0의 출력(예컨대, 302)으로 처리할 수 있다.

[0061] 또한, 제2 연산부(300)는 예측부(200)를 통해 예측된 연산픽셀(105_1~105_N)에 대응하는 커널(121)과 스트라이드를 통해, 연산픽셀(105_1~105_N)에 대한 제2 연산순서정보를 결정할 수 있다. 여기서, 제2 연산순서정보는 커널(121)과 스트라이드를 연산픽셀(105_1~105_N)에 따라 순서대로 동작시키기 위하여, 커널(121)과 입력 특징 맵(101) 간의 스트라이드에 대한 동작순서정보일 수 있다.

[0062] 그런 다음, 제2 연산부(300)는 제2 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N) 중 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 출력할 수 있다.

[0063] 실시예에 따라, 제2 연산부(300)는 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N) 중 스킵픽셀(104_1~104_N)을 0의 출력(예컨대, 302)으로 처리하는 동시에, 제2 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 출력할 수 있다. 여기서, 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)는 연산된 0의 출력(예컨대, 303_1)과 양수인 컨볼루션 정보(예컨대, 303_2)를 포함할 수 있다.

[0064] 본 출원의 실시예에 따른 합성곱 신경망의 연산 장치(10)는, 제1 연산부(100)를 통해 기설정된 모자이크 타입(106)의 예측 패턴(예컨대, 106_1~106_7)에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다. 이때, 합성곱 신경망의 연산 장치(10)는 예측부(200)를 통해 상기 적어도 하나의 픽셀정보(102_1~102_N)에 기초하여, 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)에 대해 스킵픽셀(104_1~104_N)과 연산픽셀(105_1~105_N) 중 어느 하나로 예측할 수 있다. 이후, 합성곱 신경망의 연산 장치(10)는 제2 연산부(300)를 통해 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N) 중 스킵픽셀(104_1~104_N)에 대한 합성곱 연산을 스킵하여 0의 출력(예컨대, 302)으로 처리할 수 있다. 즉, 합성곱 신경망의 연산 장치(10)는 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)에 기초하여, 보다 정확하게 스킵픽셀(104_1~104_N)을 예측할 수 있다. 이에 따라, 합성곱 신경망의 연산 장치(10)는 스킵픽셀(104_1~104_N)에 대한 합성곱 연산 동작을 스킵함으로써, 합성곱 신경망의 복잡한 연산에 따라 발생하는 시스템부하 및 전력소모를 감소시킬 수 있다.

[0065] 이하, 도 4를 참조하여, 도 1의 제1 연산부(100)에 대해 보다 구체적으로 설명될 것이다.

[0066] 도 4는 도 1의 제1 연산부(100)에 대한 실시 예이다.

[0067] 도 1 내지 도 4를 참조하면, 제1 연산부(100)는 입력 페치 회로(input fetch circuit, 110), 커널 결정 회로(kernel decision circuit, 120), 컨볼루션 회로(convolution circuit, 130), 누산회로(accumulator circuit, 140) 및 활성화 회로(activation circuit, 150) 및 풀링-서브샘플링 회로(pooling and sub-sampling circuit, 160)를 포함할 수 있다.

[0068] 보다 구체적으로, 입력 페치 회로(110)는 외부의 클라우드 서버(111) 또는 내부의 메모리(112)로부터 입력 특징 맵(101)을 검출할 수 있다. 여기서, 외부의 클라우드 서버(111) 또는 내부의 메모리(112)는 입력 특징 맵(101)을 입력받아, 입력데이터의 특징(feature)을 추출하는 중앙처리장치 또는 제어장치에 의하여 행렬의 형태인 입력 특징 맵(101)을 저장할 수 있다.

[0069] 또한, 입력 페치 회로(110)는 외부의 클라우드 서버(111) 또는 내부의 메모리(112)로부터 기설정된 모자이크 타입(106)을 검출할 수 있다. 그런 다음, 입력 페치 회로(110)는 기설정된 모자이크 타입(106)에 따라 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)을 검출할 수 있다. 이때, 예측 패턴(예컨대, 106_1~106_7)은 출력 특징 맵(301)의 픽셀행렬과 대응되는 크기를 가지는 픽셀행렬일 수 있다.

[0070] 이하, 도 5를 참조하여, 기설정된 모자이크 타입(106)에 따라 검출되는 예측 패턴(예컨대, 106_1~106_7)에 대해 보다 구체적으로 설명될 것이다.

[0071] 도 5(a)는 제1 모자이크 타입에 따른 예측 패턴(106_1)의 실시 예이다.

[0072] 도 1 내지 도 4 및 도 5(a)를 참조하면, 기설정된 모자이크 타입(106)이 제1 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_1)은 C11, C13, C15, C31, C33, C35, C51, C53, 및

C55 위치에서 픽셀정보를 가지는 행렬일 수 있다. 여기서, C는 출력 특징 맵(301)의 행렬 정보이고, N은 출력 특징 맵(301)의 행렬 정보에 대한 X축 위치정보이고, M은 출력 특징 맵(301)의 행렬 정보에 대한 Y축 위치정보일 수 있다.

- [0073] 도 5(b)는 제2 모자이크 타입에 따른 예측 패턴(106_2)의 실시 예이다.
- [0074] 도 1 내지 도 4 및 도 5(b)를 참조하면, 기설정된 모자이크 타입(106)이 제2 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_2)은 C11, C14, C22, C25, C33, C36, C41, C44, C52, C55, C63 및 C66 위치에서, 픽셀정보를 가지는 행렬일 수 있다.
- [0075] 도 5(c)는 제3 모자이크 타입에 따른 예측 패턴(106_3)의 실시 예이다.
- [0076] 도 1 내지 도 4 및 도 5(c)를 참조하면, 기설정된 모자이크 타입(106)이 제3 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_3)은 C11, C13, C15, C22, C24, C26, C31, C33, C35, C42, C44, C46, C51, C53, C55, C62, C64 및 C66 위치에서, 픽셀정보를 가지는 행렬일 수 있다.
- [0077] 도 5(d)는 제4 모자이크 타입에 따른 예측 패턴(106_4)의 실시 예이다.
- [0078] 도 1 내지 도 4 및 도 5(d)를 참조하면, 기설정된 모자이크 타입(106)이 제4 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_4)은 C11, C14, C22, C25, C33, C36, C43, C46, C52, C55, C61 및 C64 위치에서 픽셀정보를 가지는 행렬일 수 있다.
- [0079] 도 5(e)는 제5 모자이크 타입에 따른 예측 패턴(106_5)의 실시 예이다.
- [0080] 도 1 내지 도 4 및 도 5(e)를 참조하면, 기설정된 모자이크 타입(106)이 제5 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_5)은 C11, C13, C15, C22, C24, C26, C32, C34, C36, C41, C43, C45, C51, C53, C55, C62, C64 및 C66 위치에서 픽셀정보를 가지는 행렬일 수 있다.
- [0081] 도 5(f)는 제6 모자이크 타입에 따른 예측 패턴(106_6)의 실시 예이다.
- [0082] 도 1 내지 도 4 및 도 5(f)를 참조하면, 기설정된 모자이크 타입(106)이 제6 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_6)은 C11, C12, C14, C15, C22, C23, C25, C26, C31, C33, C34, C36, C41, C43, C44, C46, C52, C53, C55, C56, C61, C62, C64 및 C65 위치에서 픽셀정보를 가지는 행렬일 수 있다.
- [0083] 도 5(g)는 제7 모자이크 타입에 따른 예측 패턴(106_7)의 실시 예이다.
- [0084] 도 1 내지 도 4 및 도 5(g)를 참조하면, 기설정된 모자이크 타입(106)이 제7 모자이크 타입에 해당하고, 출력 특징 맵(301)이 6X6 픽셀 행렬(C_{NM})인 경우, 예측 패턴(106_7)은 C11, C12, C14, C15, C22, C23, C25, C26, C31, C33, C34, C36, C41, C42, C44, C45, C52, C53, C55, C56, C61, C62, C64 및 C65 위치에서 픽셀정보를 가지는 행렬일 수 있다.
- [0085] 도 1 내지 도 4를 참조하면, 커널 결정 회로(120)는 입력 패치 회로(110)를 통해 검출된 예측 패턴(예컨대, 106_1~106_7)에 대응하는 커널(121)과 스트라이드를 통해, 입력 특징 맵(101)에 대한 제1 연산순서정보를 결정할 수 있다. 예를 들면, 커널 결정 회로(120)는 예측 패턴(예컨대, 106_1~106_7)에 대응되는 입력 특징 맵(101)과 커널(121) 간의 스트라이드 순서를 제1 연산순서정보로 결정할 수 있다.
- [0086] 다음으로, 컨볼루션 회로(130)는 커널 결정 회로(120)를 통해 결정된 제1 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다.
- [0087] 다음으로, 누산회로(140)는 컨볼루션 회로(130)를 통해 출력받는 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 제1 연산순서정보에 따라 입력 특징 맵(101)에 대한 정보로 분류 또는 그룹화하고, 입력 특징 맵(101)에 대한 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 활성화 회로(150)에 출력할 수 있다. 이때, 누산회로(140)는 동일한 입력 특징 맵(101)으로부터 연산 또는 출력된 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 누적할 수 있다.
- [0088] 다음으로, 활성화 회로(150)는 누산회로(140)를 통해 출력받는 적어도 하나의 픽셀정보(102_1~102_N)를 비선형

함수를 통해 제1 연산순서정보에 따라 활성화할 수 있다. 여기서, 비선형 함수는 ReLU Function을 포함할 수 있다. 보다 구체적으로, 적어도 하나의 픽셀정보(102_1~102_N)가 음수(-)인 경우, 활성화 회로(150)는, 비선형 함수를 통해 0의 픽셀정보로 처리하고, 적어도 하나의 픽셀정보(102_1~102_N)가 양수(+)인 경우, 적어도 하나의 픽셀정보(102_1~102_N)를 그대로 출력할 수 있다.

[0089] 이때, 풀링-서브샘플링 회로(160)는 활성화 회로(150)를 통해 출력 특징 맵(301)에 대한 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다. 여기서, 적어도 하나의 픽셀정보(102_1~102_N)는 0의 픽셀과 양수의 적어도 하나의 픽셀정보(102_1~102_N)를 포함할 수 있다.

[0090] 이후, 풀링-서브샘플링 회로(160)는 출력 특징 맵(301)에 대한 적어도 하나의 픽셀정보(102_1~102_N)를 외부의 클라우드 서버(111) 또는 내부의 메모리(112)에 저장할 수 있다.

[0091] 실시예에 따라, 풀링-서브샘플링 회로(160)는 출력 특징 맵(301)에 대한 적어도 하나의 픽셀정보(102_1~102_N)를 서브샘플링하여 출력 특징 맵(301)에 대한 적어도 하나의 픽셀정보(102_1~102_N)를 작은 행렬 사이즈로 변환할 수 있다.

[0092] 이하, 도 6을 참조하여, 제2 연산부(300)에 대해 보다 구체적으로 설명될 것이다.

[0093] 도 6은 도 1의 제2 연산부(300)에 대한 실시 예이다.

[0094] 도 1 내지 도 4 및 도 7을 참조하면, 제2 연산부(300)는 입력 패치 회로(310), 커널 결정 회로(320), 컨볼루션 회로(330), 누산회로(340), 활성화 회로(350) 및 풀링-서브샘플링 회로(360)를 포함하고, 도 4에서 설명된 동일한 기능과 동작에 대한 중복된 설명은 생략한다.

[0095] 본 출원에서, 제1 및 제2 연산부(300)는 입력 패치 회로(110 또는 310), 커널 결정 회로(120 또는 320), 컨볼루션 회로(130 또는 330), 누산회로(140 또는 340), 활성화 회로(150 또는 350) 및 풀링-서브샘플링 회로(160 또는 360)를 포함하는 것으로 설명되지만, 다양한 실시예에 따라 독립된 구성으로 실시되어, 커널 결정 회로(120 또는 320), 컨볼루션 회로(130 또는 330), 누산회로(140 또는 340), 활성화 회로(150 또는 350) 및 풀링-서브샘플링 회로(160 또는 360)를 제어할 수 있다.

[0096] 먼저, 입력 패치 회로(310)는 예측부(200)를 통해 예측된 출력 특징 맵(301)의 스킵픽셀(104_1~104_N)을 0의 출력으로 모두 처리할 수 있다. 즉, 예측부(200)를 통해 예측된 출력 특징 맵(301)의 스킵픽셀(104_1~104_N)은 커널 결정 회로(320), 컨볼루션 회로(330), 누산회로(340), 활성화회로(350) 및 풀링-서브샘플링회로(360)에 대한 동작과정 없이, 0의 출력(예컨대, 302)으로 처리될 수 있다.

[0097] 다음으로, 커널 결정 회로(320)는 예측부(200)를 통해 예측된 출력 특징 맵(301)의 연산픽셀(105_1~105_N)에 대응하는 커널(121)과 스트라이드를 통해, 입력 특징 맵(101)에 대한 제2 연산순서정보를 결정할 수 있다.

[0098] 그런 다음, 컨볼루션 회로(330)는 커널 결정 회로(320)를 통해 결정된 제2 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 출력할 수 있다.

[0099] 이때, 누산회로(340)는 컨볼루션 정보(예컨대, 303)를 제2 연산순서정보에 따라 입력 특징 맵(101)에 대한 컨볼루션 정보(예컨대, 303)로 분류 또는 그룹화하고, 나머지 픽셀정보(103_1~103_N) 중 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 활성화 회로(150)에 출력할 수 있다.

[0100] 다음으로, 활성화 회로(350)는 누산회로(340)를 통해 출력받는 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 비선형 함수를 통해 제2 연산순서정보에 따라 활성화하여, 0의 출력(예컨대, 303_1) 및 양수인 컨볼루션 정보(예컨대, 303_2)를 풀링-서브샘플링 회로(360)에 전송할 수 있다.

[0101] 다음으로, 풀링-서브샘플링 회로(360)는 활성화 회로(350)를 통해 전송받는 연산픽셀(105_1~105_N)에 대한 0의 출력(예컨대, 303_1) 및 양수인 컨볼루션 정보(예컨대, 303_2)와 입력 패치 회로(310)를 통해 처리된 스킵픽셀(104_1~104_N)에 대한 0의 출력(예컨대, 302)을 누적하여, 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)를 출력할 수 있다. 그런 다음, 풀링-서브샘플링 회로(360)는 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)를 외부의 클라우드 서버(111) 또는 내부의 메모리(112)에 저장할 수 있다.

[0102] 실시예에 따라, 풀링-서브샘플링 회로(360)는 병합회로(361)를 더 포함할 수 있다.

[0103] 여기서, 병합회로(361)는 제1 연산부(100)의 풀링-서브샘플링 회로(160)를 통해 출력 특징 맵(301)의 적어도 하

나의 픽셀정보(102_1~102_N)를 출력받고, 제2 연산부(300)의 풀링-서브샘플링 회로(360)를 통해 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)를 출력받을 수 있다. 이때, 병합회로(361)는 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)와 나머지 픽셀정보(103_1~103_N)를 병합하여, 출력 특징 맵(301)을 외부의 클라우드 서버(111) 또는 내부의 메모리(112)에 저장할 수 있다.

[0104] 실시예에 따라, 병합회로(361)는 외부의 클라우드 서버(111) 또는 내부의 메모리(112)에 저장된 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 검출하고, 적어도 하나의 픽셀정보(102_1~102_N)와 나머지 픽셀정보(103_1~103_N)를 검출할 수 있다.

[0105] 이에, 제2 연산부(300)는 입력 패치 회로(310)를 통해 출력 특징 맵(301)의 스킵픽셀을 0의 출력(예컨대, 302)으로 처리할 때, 커널 결정 회로(320), 컨볼루션 회로(330), 누산회로(340), 활성화 회로(350) 및 풀링-서브샘플링 회로(360)를 통해 출력 특징 맵(301)의 연산픽셀(105_1~105_N)을 병렬로 처리하여, 출력 특징 맵(301)을 출력할 수 있다.

[0106] 도 7은 본 출원의 다른 실시예에 따르는 합성곱 신경망의 연산 장치(11)의 블록도이다.

[0107] 도 7을 참조하면, 합성곱 신경망의 연산 장치(11)는 제1 연산부(100), 예측부(200), 제2 연산부(300) 및 희소 행렬 가속부(Sparse Matrix Accelerator, 400)를 포함할 수 있다. 이하에서, 제1 연산부(100), 예측부(200) 및 제2 연산부(300)는 도 1 내지 도 10에서 설명된 기능, 구성 및 동작이 동일하므로, 중복된 설명은 생략한다.

[0108] 먼저, 희소 행렬 가속부(400)는 제1 및 제2 연산부(100, 300)를 통해 기설정된 모자이크 타입(106)에 따라 출력된 출력 특징 맵(301)을 저장할 수 있다. 이때, 희소 행렬 가속부(400)는 출력 특징 맵(301)에서 0의 출력을 제외한 픽셀정보에 대응하는 커널(121)과 스트라이드를 통해, 제3 연산순서정보를 결정할 수 있다. 여기서, 제3 연산순서정보는 커널(121)과 스트라이드를 출력 특징 맵(301)에서 0의 출력이 아닌 픽셀정보에 따라 순서대로 가속하여 동작시키기 위한 스트라이드 크기에 따른 순서정보일 수 있다.

[0109] 그런 다음, 희소 행렬 가속부(400)는 제3 연산순서정보에 따라, 클라우드 서버(111) 또는 메모리(112)에 새로 생성되는 다음의 입력 특징 맵(101_1)을 합성곱 연산할 수 있다. 즉, 희소 행렬 가속부(400)는 제3 연산순서정보에 따라, 새로 생성되는 다음의 입력 특징 맵(101_1)과 커널(121)을 합성곱 연산하여, 출력 특징 맵(301_1)을 보다 빠르게 출력할 수 있다.

[0110] 도 8은 합성곱 신경망의 컨볼루션 시스템(1000)에 대한 실시 예이다.

[0111] 도 1 내지 도 8을 참조하면, 합성곱 신경망의 컨볼루션 시스템(1000)은 합성곱 신경망의 연산 장치(10), 통신장치(20), 중앙처리장치(30), 저장장치(40) 및 출력장치(50)를 포함할 수 있다.

[0112] 먼저, 합성곱 신경망의 연산 장치(10)는 하드웨어 칩 형태로 장착되는 그래픽스 처리 장치(Graphics Processing Unit)로 구현되어, 도 1 내지 도 7에서 설명된 기능과 동작을 수행할 수 있다. 예를 들면, 합성곱 신경망의 연산 장치(10)는 중앙처리장치(CPU), 마이크로프로세서, 주문형 반도체(Application Specific Integrated Circuit, ASIC), Field Programmable Gate Arrays(FPGA) 등의 프로세서로 구현될 수 있다. 또한, 합성곱 신경망의 연산 장치(10)는 PIM(Processor In Memory)와 같은 지능형 반도체로 구현될 수도 있다.

[0113] 다음으로, 통신장치(20)는 네트워크(25)를 통해 입력 특징 맵(101)을 전송받고, 출력 특징 맵(301)을 전송할 수 있다. 예를 들면, 통신장치(20)는 LTE, LTE-A(LTE Advance), CDMA(code division multiple access), WCDMA(wideband CDMA), UMTS(universal mobile telecommunications system), WiBro(Wireless Broadband), 또는 GSM(Global System for Mobile Communications) 등 중 적어도 하나의 무선통신을 통해 입력 특징 맵(101)을 전송받고, 출력 특징 맵(301)을 전송할 수 있다. 또한, 통신장치(20)는 WiFi(wireless fidelity), 블루투스, 블루투스 저전력(BLE), 지그비(Zigbee), NFC(near field communication), 자력 시큐어 트랜스미션(Magnetic Secure Transmission), 라디오 프리퀀시(RF), 또는 보디 에어리어 네트워크(BAN) 중 적어도 하나의 근거리 통신을 통해 입력 특징 맵(101)을 전송받고, 출력 특징 맵(301)을 전송받을 수 있다. 또한, 통신장치(20)는 USB(universal serial bus), HDMI(high definition multimedia interface), RS-232(recommended standard232), 전력선 통신, 또는 POTS(plain old telephone service) 등 중 적어도 하나의 유선통신을 통해 입력 특징 맵(101)을 전송받고, 출력 특징 맵(301)을 전송받을 수 있다.

[0114] 여기서, 네트워크(25)는 TCP/IP 프로토콜 및 그 상위계층에 존재하는 여러 서비스, 즉 HTTP(Hyper Text Transfer Protocol), Telnet, FTP(File Transfer Protocol), DNS(Domain Name System), SMTP(Simple Mail Transfer Protocol), SNMP(Simple Network Management Protocol), NFS(Network File Service), NIS(Network

Information Service) 등을 제공하는 전 세계적인 개방형 컴퓨터 네트워크 구조를 의미할 수 있다.

- [0115] 실시예에 따라, 합성곱 신경망의 컨볼루션 시스템(1000)은 인터페이스부(60)를 더 포함하고, 인터페이스부(60)는 외부장치로부터 입력데이터에 해당하는 입력 특징 맵(101)을 입력받는 USB, UART 또는 SD카드 중 적어도 하나를 포함할 수 있다.
- [0116] 다음으로, 중앙처리장치(30)는 통신장치(20)를 통해 전송받는 또는 인터페이스부(60)를 통해 입력 특징 맵(101)을 입력받을 수 있다. 이때, 합성곱 신경망의 연산 장치(10)는 기설정된 모자이크 타입(106)에 기초하여, 입력 특징 맵(101)을 합성곱 연산하여 출력 특징 맵(301)을 출력할 수 있다.
- [0117] 그런 다음, 중앙처리장치(30)는 합성곱 신경망의 연산 장치(10)를 통해 출력된 출력 특징 맵(301)을 출력할 수 있다. 이때, 중앙처리장치(30)는, 출력 특징 맵(301)을 저장장치(40)에 저장하거나, 통신장치(20)를 통해 출력 특징 맵(301)을 전송하거나 또는, 출력장치(50)를 통해 출력 특징 맵(301)을 표시할 수 있다.
- [0118] 즉, 합성곱 신경망의 컨볼루션 시스템(1000)은 네트워크(25)를 통해 통신장치(20) 또는 인터페이스부(60)로부터 입력 특징 맵(101)을 전송 또는 입력받아, 출력 특징 맵(301)을 전송할 수 있는 컴퓨터, 휴대용 단말기, 텔레비전 또는 클라우드 서버로 구현될 수 있다.
- [0119] 예를 들면, 컴퓨터는 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop) 등을 포함하고, 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, PCS(Personal Communication System), GSM(Global System for Mobile communications), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access) 2000, W-CDMA(W-Code Division Multiple Access), Wibro(Wireless Broadband Internet) 단말, 스마트폰(Smart Phone) 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치를 포함할 수 있다. 또한, 텔레비전은 IPTV(Internet Protocol Television), 인터넷 TV(Internet Television), 지상파 TV, 케이블 TV 등을 포함할 수 있다.
- [0120] 또한, 클라우드 서버는 클라우드 컴퓨팅, 분산 컴퓨팅, 또는 클라이언트-서버 컴퓨팅 기술을 이용하여, 입력데이터별로 출력 특징 맵(301)에 따른 출력데이터를 데이터베이스(DB)화하여 저장, 분류, 관리 및 제공할 수 있다. 이때, 클라우드 서버는 저장DB를 포함하고, 이때, 저장DB는 오라클(Oracle), 인포믹스(Infomix), 사이베이스(Sybase), DB2와 같은 관계형 데이터베이스 관리 시스템(RDBMS)이나, 겜스톤(Gemston), 오리온(Orion), O2 등과 같은 객체 지향 데이터베이스 관리 시스템(OODBMS)을 이용하여 본 발명의 목적에 맞게 구현될 수 있고, 자신의 기능을 달성하기 위하여 적당한 필드(field)들을 가지고 있다.
- [0121] 도 9는 본 출원의 실시예에 따른 합성곱 신경망의 연산 장치(10)의 동작 프로세스이다.
- [0122] 도 1 내지 도 9를 참조하면, S110 단계에서, 제1 연산부(100)는 기설정된 모자이크 타입(106)에 기초하여, 입력 특징 맵(101)을 합성곱 연산하고, 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다.
- [0123] 다음으로, S120 단계에서, 예측부(200)는 제1 연산부(100)를 통해 출력된 적어도 하나의 픽셀정보(102_1~102_N)에 기초하여, 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)에 대해 스킵픽셀(104_1~104_N)과 연산픽셀(105_1~105_N) 중 어느 하나로 예측할 수 있다.
- [0124] 다음으로, S130 단계에서, 제2 연산부(300)는 스킵픽셀(104_1~104_N)에 대한 합성곱 연산을 스킵하여 0의 출력(예컨대, 302)으로 처리할 수 있다.
- [0125] 이때, S140 단계에서, 제2 연산부(300)는 연산픽셀(105_1~105_N)에 대응하는 커널(121)과 스트라이드를 통해, 입력 특징 맵(101)을 합성곱 연산하여 나머지 픽셀정보(103_1~103_N)의 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 출력할 수 있다.
- [0126] 이후, S150 단계에서, 제2 연산부(300)는 0의 출력(예컨대, 302)과 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 병합한 나머지 픽셀정보(103_1~103_N)와 제1 연산부(100)를 통해 출력된 적어도 하나의 픽셀정보(102_1~102_N)를 병합하여 출력 특징 맵(301)을 출력할 수 있다.
- [0127] 도 10은 도 1의 제1 연산부(100)에 대한 동작 프로세스이다.
- [0128] 도 1 내지 도 10을 참조하면, 먼저, S210 단계에서, 제1 연산부(100)는 기설정된 모자이크 타입(106)에 기초하여, 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)을 검출할 수 있다. 예를 들면, 제1 연산

부(100)는 제1 모자이크 타입에 기초하여, 미리 정해진 출력 특징 맵(301)의 예측 패턴(예컨대, 106_1~106_7)을 검출할 수 있다.

[0129] 그런 다음, S220 단계에서, 제1 연산부(100)는 예측 패턴(예컨대, 106_1~106_7)에 대응하는 커널(121)과 스트라이드를 통해, 입력 특징 맵(101)에 대한 제1 연산순서정보를 결정할 수 있다.

[0130] 이후, S230 단계에서, 제1 연산부(100)는 제1 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여, 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)를 출력할 수 있다.

[0131] 도 11은 도 1의 예측부(200)에 대한 동작 프로세스이다.

[0132] 도 1 내지 도 11을 참조하면, S310 단계에서, 예측부(200)는 제1 연산부(100)를 통해 적어도 하나의 픽셀정보(102_1~102_N)와 예측 패턴(예컨대, 106_1~106_7)을 전송받을 수 있다.

[0133] 그런 다음, S320 단계에서, 예측부(200)는 적어도 하나의 픽셀정보(102_1~102_N)와 예측 패턴(예컨대, 106_1~106_7)을 비교할 수 있다.

[0134] 이후, S330 단계에서, 예측부(200)는 적어도 하나의 픽셀정보(102_1~102_N)의 0의 출력과 예측 패턴(예컨대, 106_1~106_7) 간의 동일 여부에 따라, 나머지 픽셀정보(103_1~103_N)에 기설정된 제1 영역(104)을 스킵픽셀(104_1~104_N)로 예측하고, 나머지 영역(105)을 연산픽셀(105_1~105_N)로 예측할 수 있다.

[0135] 즉, 예측부(200)는 적어도 하나의 픽셀정보(102_1~102_N)의 0의 출력과 예측 패턴(예컨대, 106_1~106_7)의 0의 출력을 비교하여, 동일 여부에 따라, 나머지 픽셀정보(103_1~103_N)에 대한 제1 영역(104)과 나머지 영역(105)으로 식별할 수 있다.

[0136] 도 12는 도 1의 제2 연산부(300)에 대한 동작 프로세스이다.

[0137] 도 1 내지 도 12를 참조하면, 먼저, S410 단계에서, 제2 연산부(300)는 나머지 픽셀정보(103_1~103_N) 중 예측부(300)를 통해 예측된 스킵픽셀(104_1~104_N)을 0의 출력(예컨대, 302)으로 처리할 수 있다.

[0138] 그런 다음, S420 단계에서, 제2 연산부(300)는 나머지 픽셀정보(103_1~103_N) 중 예측부(300)를 통해 예측된 연산픽셀(105_1~105_N)에 대응하는 커널(121)과 스트라이드를 통해 입력 특징 맵(101)에 대한 제2 연산순서정보를 결정할 수 있다.

[0139] 그런 다음, S430 단계에서, 제2 연산부(300)는 제2 연산순서정보에 따라, 입력 특징 맵(101)과 커널(121)을 합성곱 연산하여 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 출력할 수 있다.

[0140] 이때, S440 단계에서, 제2 연산부(300)는 0의 출력(예컨대, 302)과 연산픽셀(105_1~105_N)에 대한 컨볼루션 정보(예컨대, 303)를 병합하여 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)를 출력할 수 있다.

[0141] 이후, S450 단계에서, 제2 연산부(300)는 제1 연산부(100)를 통해 출력된 출력 특징 맵(301)의 적어도 하나의 픽셀정보(102_1~102_N)와 출력 특징 맵(301)의 나머지 픽셀정보(103_1~103_N)를 병합하여, 출력 특징 맵(301)을 출력할 수 있다.

[0142] 본 출원은 도면에 도시된 일 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 출원의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0143] 10, 11: 합성곱 신경망의 연산 장치

100: 제1 연산부

110, 310: 입력 패치 회로

120, 320: 커널 결정 회로

130, 330: 컨볼루션 회로

140, 340: 누산회로

150, 350: 활성화 회로

160, 360: 풀링-서브샘플링 회로

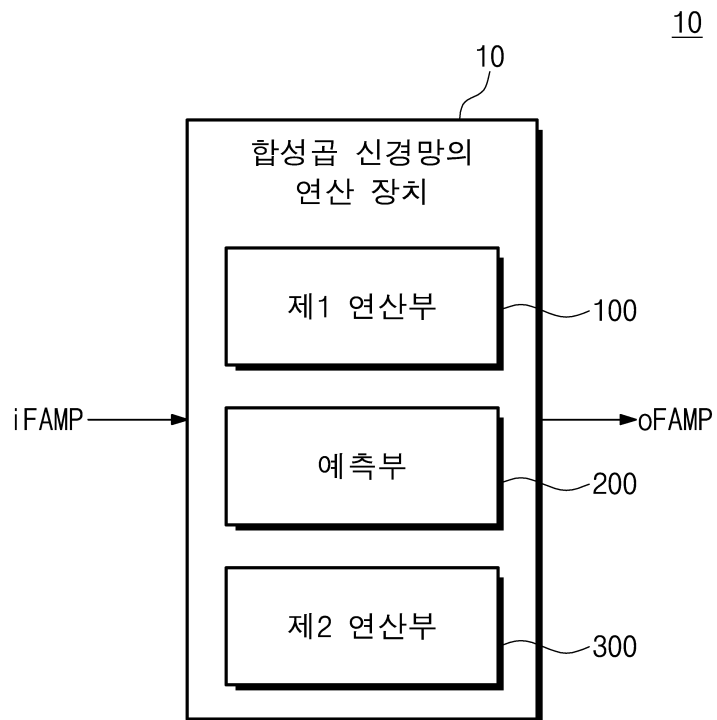
200: 예측부

300: 제2 연산부

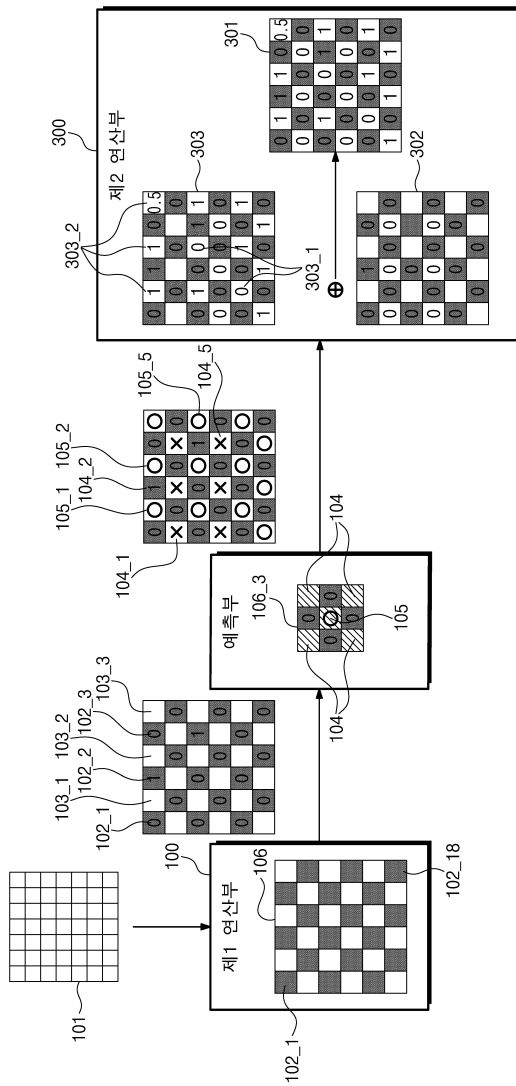
1000: 합성곱 신경망의 컨볼루션 시스템

도면

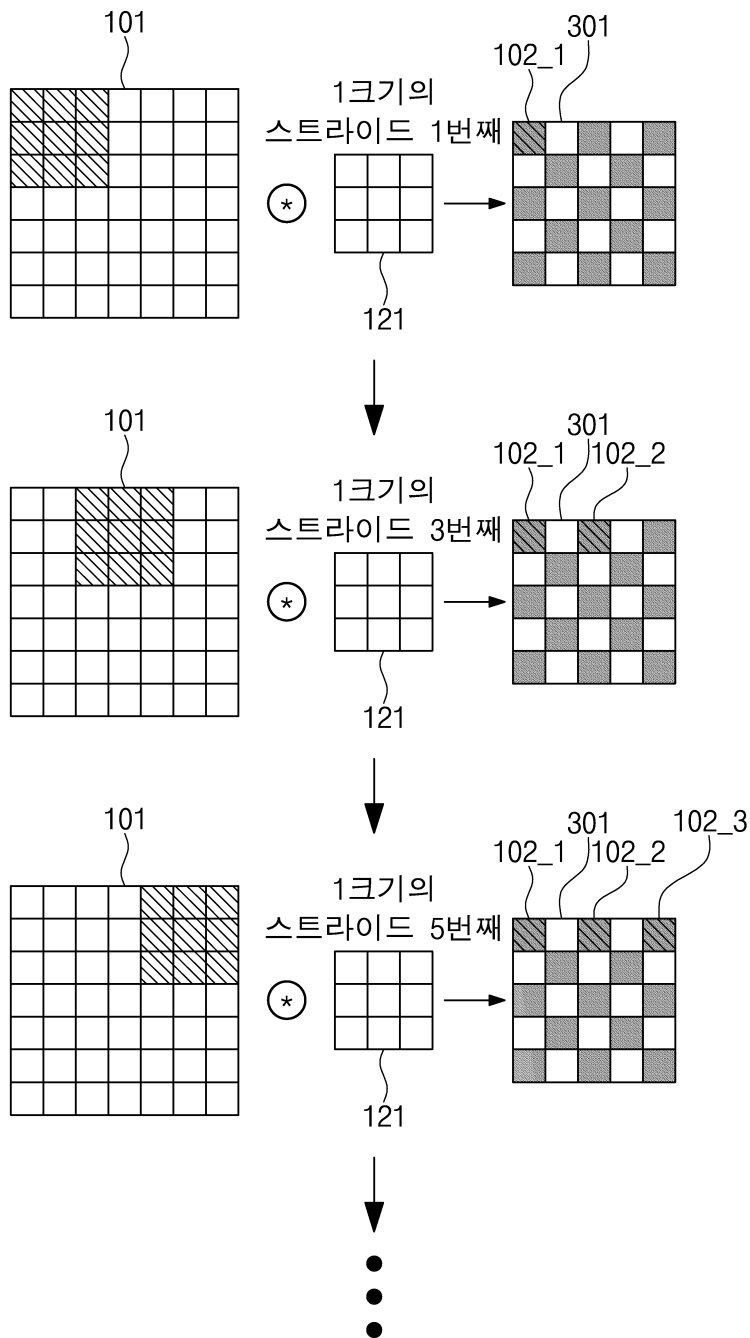
도면1



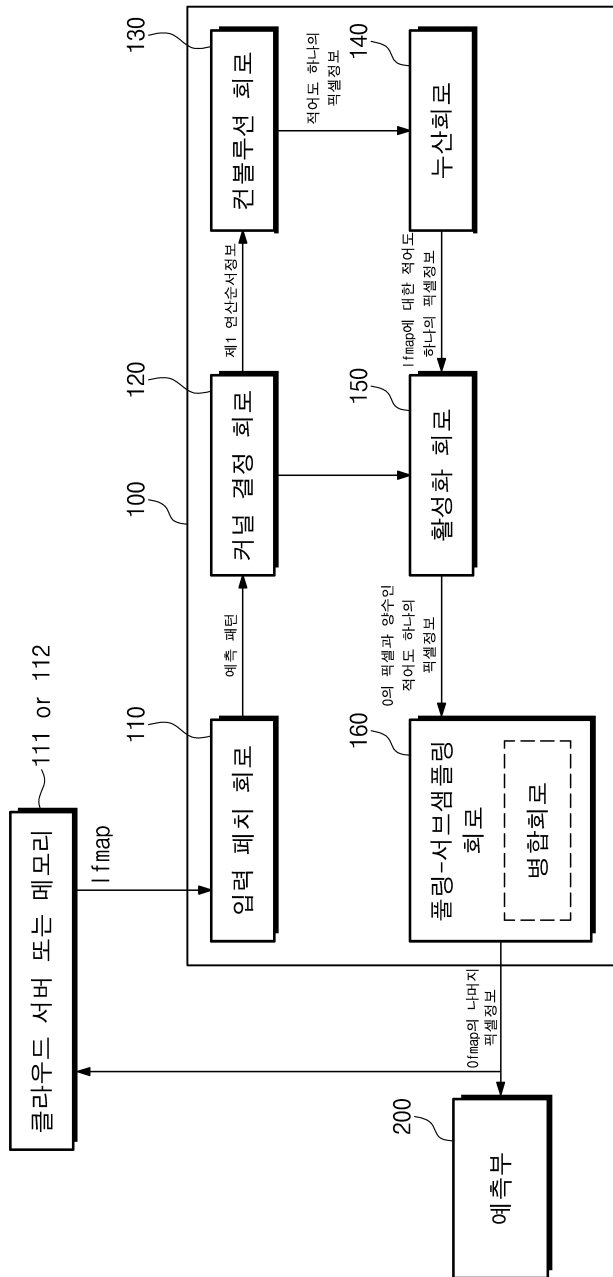
도면2



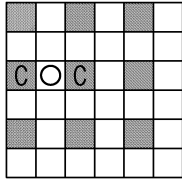
도면3



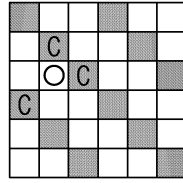
도면4



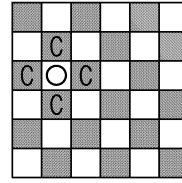
도면5



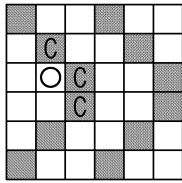
(a)



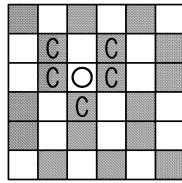
(b)



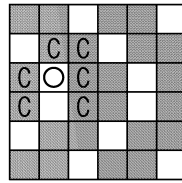
(c)



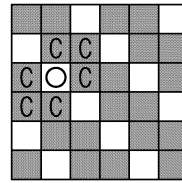
(d)



(e)

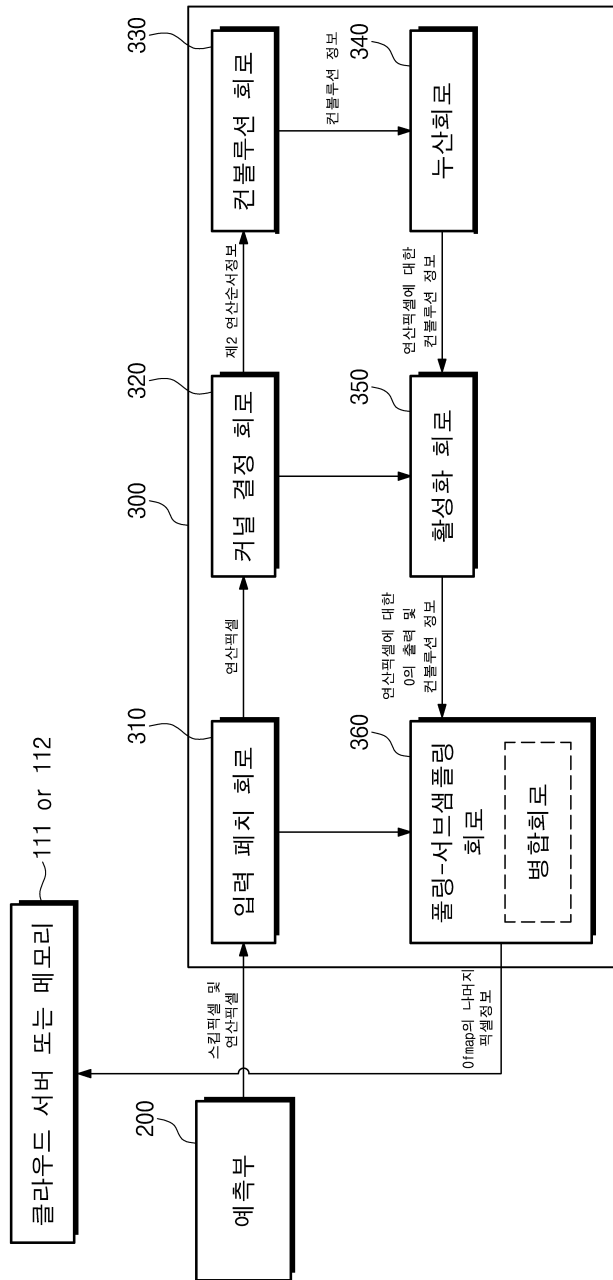


(f)

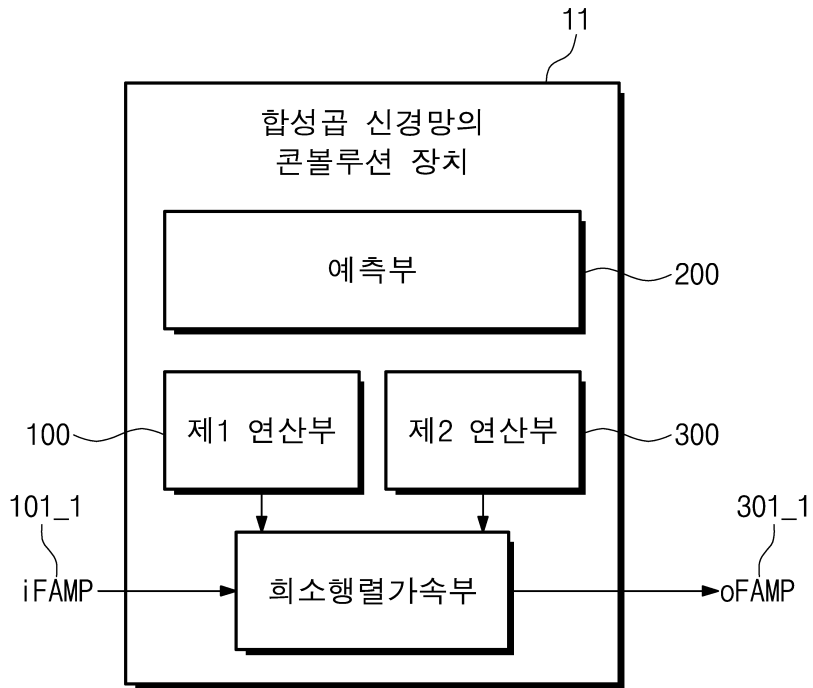


(g)

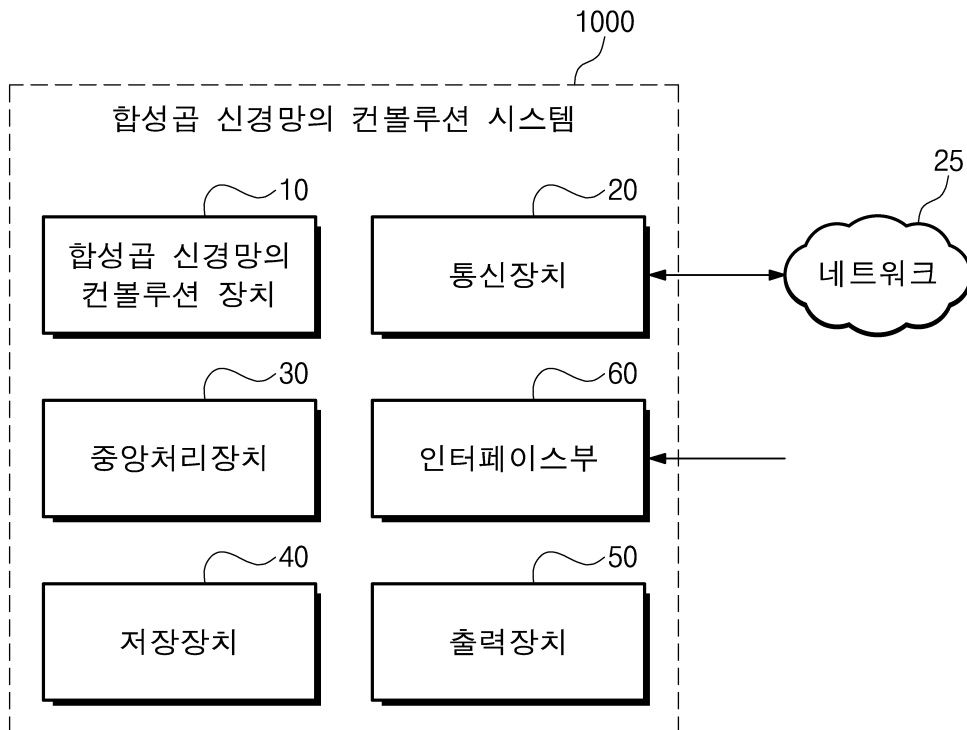
도면6



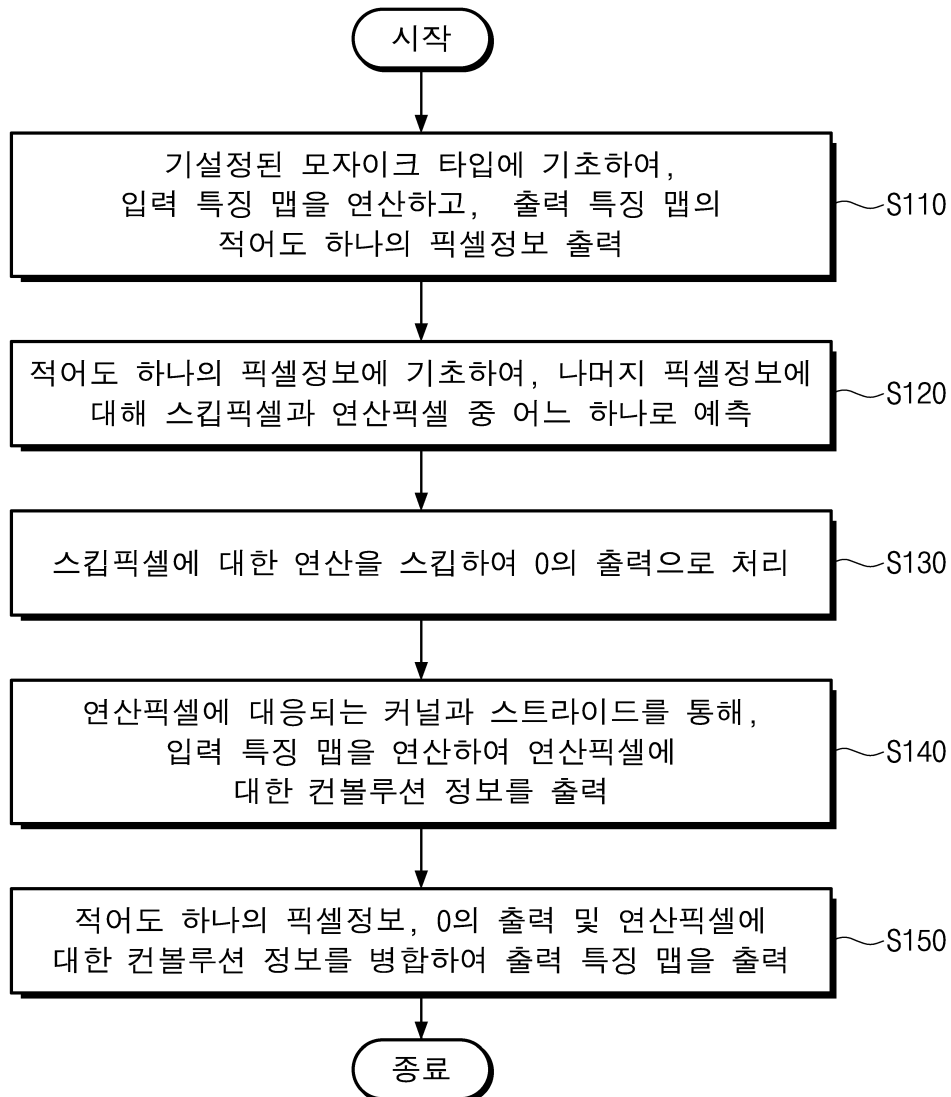
도면7



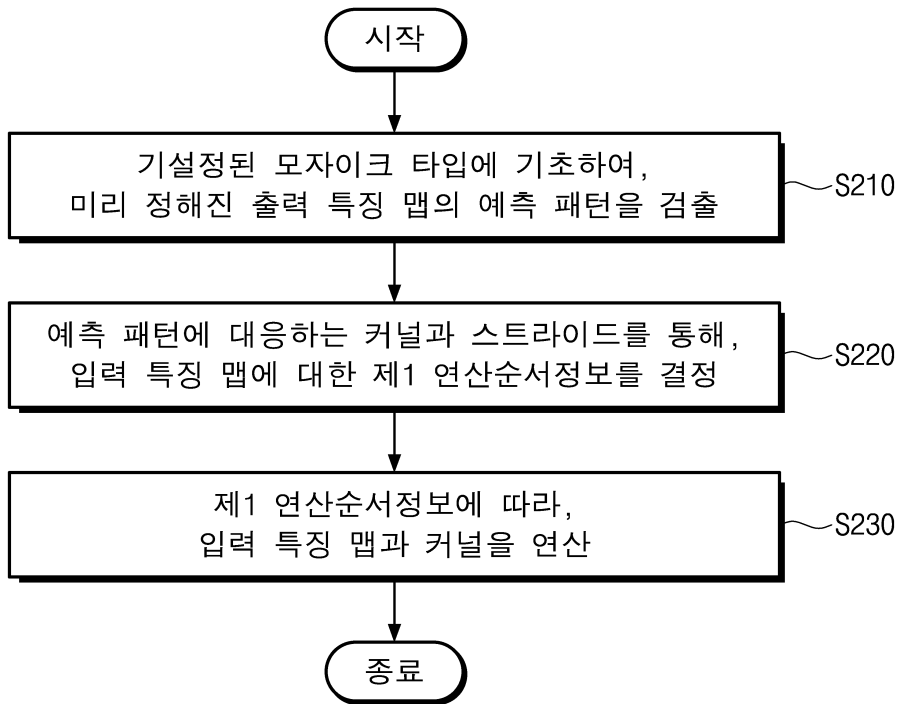
도면8



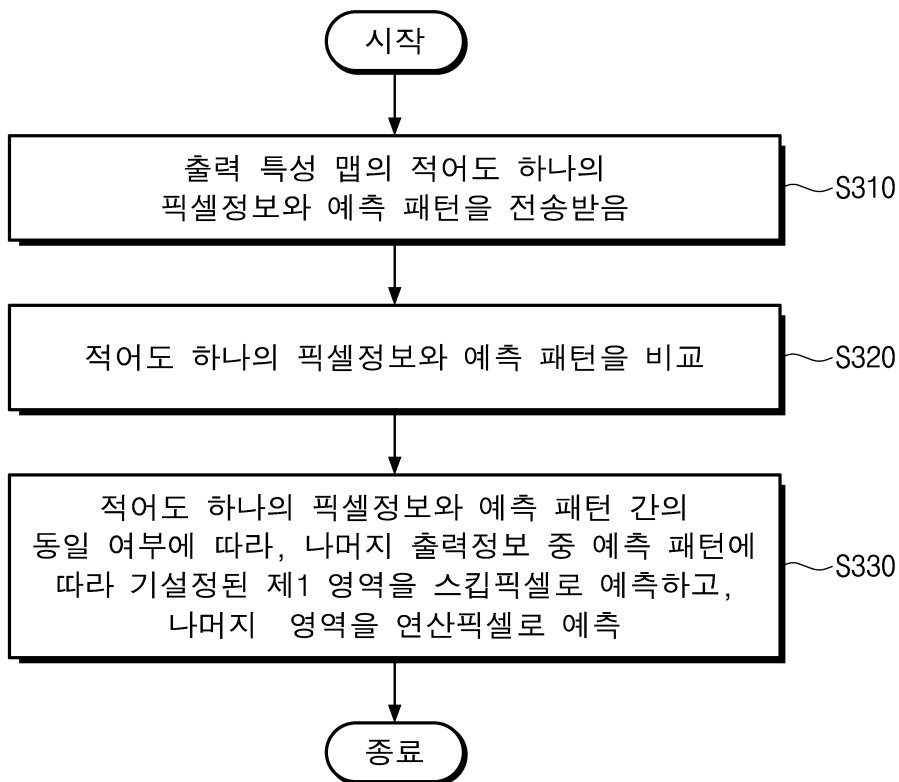
도면9



도면10



도면11



도면12

