(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년03월28일
(11) 등록번호 10-2652476
(24) 등록일자 2024년03월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/063 (2023.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/063 (2013.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2021-0103968
(22) 출원일자 2021년08월06일
심사청구일자 2021년08월06일
(65) 공개번호 10-2022-0097161
(43) 공개일자 2022년07월07일
(30) 우선권주장
1020200189766 2020년12월31일 대한민국(KR)
(56) 선행기술조사문헌
KR1020190129240 A*
US20180314941 A1*
KR1020210014056 A
US10699189 B2
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
주식회사 딥엑스
경기도 성남시 분당구 판교역로241번길 20, 5층 (삼평동, 미래에셋벤처타워)
(72) 발명자
김녹원
경기도 성남시 분당구 판교역로 231 에이치스퀘어 에스동 7층 701호 (주)딥엑스
(74) 대리인
특허법인인벤싱크

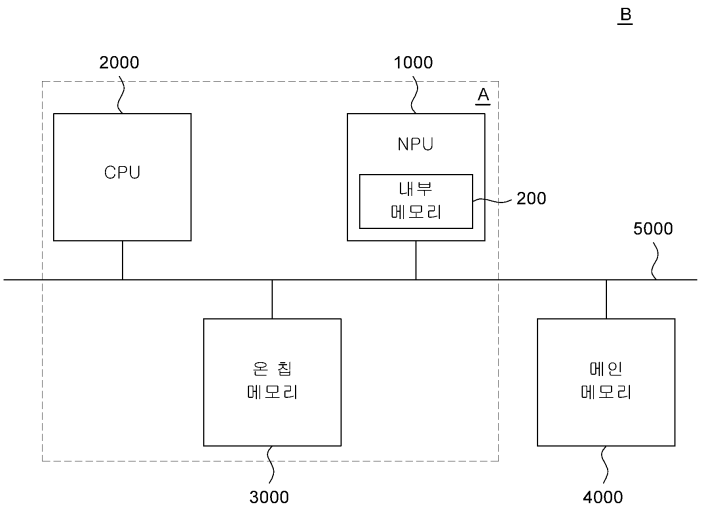
전체 청구항 수 : 총 32 항 심사관 : 이준상

(54) 발명의 명칭 **인공신경망을 위한 방법 및 신경 프로세싱 유닛**

(57) 요 약

본 발명의 일 예시에 따른 인공신경망을 위한 방법이 제공된다. 상기 방법은 ANN (artificial neural network)에 대한 동작들을 수행하는 단계를 포함하고, 동작들을 위해, 복수의 배치채널들은 제 1 배치채널 및 제 2 배치채널을 포함하고, 동작들은, 적어도 하나의 메모리, 일 세트의 가중치, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부를 저장하는 단계; 그리고 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부 각각과 일 세트의 가중치 값들을 계산하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126457
과제번호	2020-0-00364-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능반도체응용기술개발(R&D)
연구과제명	AI 기반 차량용 통신 기술 향상을 위한 NPU와 응용 시스템 개발
기 여 율	1/1
과제수행기관명	(주)딥엑스
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

ANN (artificial neural network)에 대한 동작들을 수행하는 단계를 포함하고,

상기 동작들을 위해, 복수의 배치채널들은 제 1 배치채널 및 제 2 배치채널을 포함하고,

상기 동작들은,

일 세트의 가중치, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부를 적어도 하나의 메모리에 포함된 온-칩 메모리 및/또는 내부 메모리에 저장하는 단계;

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치를 계산하는 단계; 및

상기 복수의 배치채널들 각각의 적어도 일부가 계산이 완료될 때까지 상기 일 세트의 가중치를 상기 온-칩 메모리 및/또는 내부 메모리에 유지하면서 다음 계산될 복수의 채널들을 로드하는 단계를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 2

제 1 항에 있어서,

상기 제 1 배치채널의 적어도 일부의 크기 및 상기 제 2 배치채널의 적어도 일부의 크기는 실질적으로 동일한,

신경 프로세싱 유닛의 동작 방법.

청구항 3

제 1 항에 있어서,

상기 동작들을 위해, 상기 일 세트의 가중치는 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부에 대응하는,

신경 프로세싱 유닛의 동작 방법.

청구항 4

제 1 항에 있어서,

상기 방법은, 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치를 계산하는 동안, 상기 적어도 하나의 메모리의 적어도 일부에 다음에 계산될 상기 제 1 배치채널의 다른 적어도 일부를 저장하는 단계를 더 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 5

제 1 항에 있어서,

상기 동작들을 위해, 상기 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 더 포함하고,

상기 동작들은,

상기 일 세트의 가중치를 유지하면서 상기 제 3 배치채널의 적어도 일부 및 상기 제 4 배치채널의 적어도 일부를 상기 적어도 하나의 메모리에 저장하는 단계 및

상기 제 3 배치채널의 적어도 일부 및 상기 제 4 배치채널의 적어도 일부와 상기 일 세트의 가중치를 계산하는

단계를 더 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 6

삭제

청구항 7

제 1 항에 있어서,

상기 동작들은 다음 세트의 가중치, 상기 제 1 배치채널의 다음 부분 및 상기 제 2 배치채널의 다음 부분을 상기 온-칩 메모리 및/또는 내부 메모리에 저장하는 단계; 및

상기 제 1 배치채널의 다음 부분 및 상기 제 2 배치채널의 다음 부분과 상기 다음 세트의 가중치를 계산하는 단계를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 8

제 1 항에 있어서,

상기 동작들은

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치로부터 계산된 제 1 값들을 상기 적어도 하나의 메모리에 저장하는 단계;

다음 프로세싱 단계를 위해 상기 적어도 하나의 메모리에 다음 세트의 가중치를 저장하는 단계; 그리고

상기 제 1 값들과 상기 다음 세트의 가중치를 계산하는 단계를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 9

제 8 항에 있어서,

상기 적어도 하나의 메모리는 내부 메모리를 포함하고,

상기 제 1 값들과 상기 다음 세트의 가중치를 계산한 제 2 값들은 상기 내부 메모리에서만 상주하는,

신경 프로세싱 유닛의 동작 방법.

청구항 10

삭제

청구항 11

제 1 항에 있어서,

상기 동작들은, 상기 내부 메모리에 저장하기 전에, 상기 일 세트의 가중치의 크기, 상기 제 1 배치채널의 적어도 일부의 크기 및 상기 제 2 배치채널의 적어도 일부의 크기를 상기 내부 메모리에 피팅되도록 타일링하는 단계를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 12

제 1 항에 있어서,

상기 동작들에 대해,

상기 ANN은 상기 복수의 배치채널들로부터의 객체 검출, 분류 또는 세그먼트화를 포함하는 적어도 하나의 동작

을 수행하도록 구성되는,

신경 프로세싱 유닛의 동작 방법.

청구항 13

제 12 항에 있어서,

상기 동작들을 위해,

상기 객체는 차량, 신호등, 장애물, 통행인, 사람, 동물, 도로, 표지판, 및 도로 선 중 적어도 하나를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 14

제 1 항에 있어서,

상기 ANN에 대한 동작들 전에 상기 복수의 배치채널들을 전처리하는 단계를 포함하는, 신경 프로세싱 유닛의 동작 방법.

청구항 15

제 12 항에 있어서,

상기 동작들을 위해,

상기 ANN은 객체들의 향상된 검출을 위해 상기 복수의 배치채널들을 전처리하면서, 상기 복수의 배치채널들로부터 상기 객체를 동시에 검출하도록 구성되는,

신경 프로세싱 유닛의 동작 방법.

청구항 16

제 1 항에 있어서,

상기 동작들에 대해, 상기 복수의 배치채널들 각각은 복수의 이미지들 각각에 대응하는, 신경 프로세싱 유닛의 동작 방법.

청구항 17

제 16 항에 있어서,

상기 동작들을 위해, 상기 복수의 배치채널들 중 적어도 하나의 배치채널은 IR, RGB, YCBCR, HSV, 및 HIS 형식 중 하나의 형식인,

신경 프로세싱 유닛의 동작 방법.

청구항 18

제 16 항에 있어서,

상기 동작들을 위해, 상기 복수의 배치채널들 중 적어도 하나는 차량의 내부를 캡처하는 이미지를 포함하고, 상기 ANN은 차량의 안전 관련 객체, 기능, 운전자의 상태, 및 승객의 상태 중 적어도 하나를 검출하도록 구성되는,

신경 프로세싱 유닛의 동작 방법.

청구항 19

제 16 항에 있어서,

상기 동작들을 위해, 상기 복수의 이미지들은 RGB 이미지, IR 이미지, RADAR 이미지, ULTRASOUND 이미지, LIDAR 이미지, 열 화상 이미지, NIR 이미지 및 이들의 융합 이미지 중 적어도 하나를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 20

제 16 항에 있어서,

상기 동작들을 위해, 상기 복수의 이미지들은 실질적으로 동일한 기간에 캡처된 이미지들인,

신경 프로세싱 유닛의 동작 방법.

청구항 21

제 1 항에 있어서,

상기 동작들을 위해, 상기 복수의 배치채널들 각각은 복수의 센서 데이터 각각에 대응하고, 그리고 상기 복수의 센서 데이터는 압력 센서, 피에조 센서, 습도 센서, 먼지 센서, 스모그 센서, 소나 센서, 진동 센서, 가속도 센서 또는 모션 센서 중 하나 이상으로부터의 데이터를 포함하는,

신경 프로세싱 유닛의 동작 방법.

청구항 22

제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크에 대한 신경 프로세싱 유닛으로서,

온-칩 메모리 및/또는 내부 메모리를 포함하며, 상기 제 1 배치채널의 적어도 일부, 상기 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치를 저장하도록 구성된 적어도 하나의 내부 메모리;

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치를 계산하되, 상기 복수의 배치채널들 각각의 적어도 일부가 계산이 완료될 때까지 상기 일 세트의 가중치를 상기 온-칩 메모리 및/또는 내부 메모리에 유지하면서 다음 계산될 복수의 채널들을 로드하는 스케줄러; 및

상기 일 세트의 가중치를 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element)를 포함하는, 신경 프로세싱 유닛.

청구항 23

제 22 항에 있어서,

상기 적어도 하나의 내부 메모리에 할당되는 상기 제 1 배치채널의 적어도 일부의 크기 및 상기 제 2 배치채널의 적어도 일부의 크기는 실질적으로 동일한, 신경 프로세싱 유닛.

청구항 24

제 22 항에 있어서,

상기 일 세트의 가중치는 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부에 대응하는, 신경 프로세싱 유닛.

청구항 25

제 22 항에 있어서,

상기 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 포함하고,

상기 적어도 하나의 내부 메모리는, 상기 일 세트의 가중치를 유지하는 동안, 상기 제 3 배치채널의 적어도 일부 및 상기 제 4 배치채널의 적어도 일부를 저장하도록 더 구성되고,

상기 PE는 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부 각각과 상기 일 세트의 가중치를 계산하도록 더 구성되는, 신경 프로세싱 유닛.

청구항 26

삭제

청구항 27

제 22 항에 있어서,

상기 적어도 하나의 내부 메모리는 또 다른 세트의 가중치, 상기 제 1 배치채널의 다음 부분 및 상기 제 2 배치채널의 다음 부분을 저장하도록 더 구성되고; 그리고

상기 PE는 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분 각각과 상기 또 다른 세트의 가중치를 계산하도록 더 구성되는, 신경 프로세싱 유닛.

청구항 28

제 22 항에 있어서,

상기 적어도 하나의 내부 메모리는 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치로부터 계산된 값들을 저장하고, 그리고 다음 스테이지에 대한 일 세트의 가중치를 저장하도록 더 구성되고,

상기 PE는 상기 계산된 값들과 상기 다음 스테이지에 대한 일 세트의 가중치를 계산하도록 더 구성되고,

상기 일 세트의 가중치는 상기 복수의 배치채널들이 계산될 때까지 상기 내부 메모리에 유지되는, 신경 프로세싱 유닛.

청구항 29

제 22 항에 있어서,

상기 적어도 하나의 내부 메모리는,

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부에 대응하고,

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치 로부터 계산된 제 1 계산 값들을 저장하고,

다음 프로세싱 단계를 위한 다음 세트의 가중치를 저장하도록 더 구성되고,

상기 PE는 상기 제 1 계산 값들과 상기 다음 세트의 가중치를 계산하도록 더 구성되는, 신경 프로세싱 유닛.

청구항 30

제 22 항에 있어서,

상기 스케줄러는 상기 일 세트의 가중치의 크기, 상기 제 1 배치채널의 적어도 일부의 크기 및 상기 제 2 배치채널의 적어도 일부의 크기를 상기 내부 메모리에 맞게 조정하도록 구성되는, 신경 프로세싱 유닛.

청구항 31

제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크 (ANN) 를 위한 신경 프로세싱 유닛으로서,

온-칩 메모리 및/또는 내부 메모리를 포함하며, 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치를 저장하도록 구성된 적어도 하나의 내부 메모리;

상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치를 계산하되, 상기 복수의 배치채널들 각각의 적어도 일부가 계산이 완료될 때까지 상기 일 세트의 가중치를 상기 온-칩 메모리 및/또는 내부 메모리에 유지하면서 다음 계산될 복수의 채널들을 로드하는 스케줄러; 및

상기 일 세트의 가중치를 상기 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element) 를 포함하고,

상기 제 1 배치채널의 적어도 일부의 크기는, 상기 적어도 하나의 내부 메모리의 크기를 상기 복수의 배치채널

들의 수로 나눈 것과 같거나 작은, 신경 프로세싱 유닛.

청구항 32

제 31 항에 있어서,

상기 적어도 하나의 내부 메모리의 크기는 상기 ANN의 가장 큰 특징맵의 크기 및 상기 배치채널들의 수에 대응하는, 신경 프로세싱 유닛.

청구항 33

제 31 항에 있어서,

상기 적어도 하나의 내부 메모리는 상기 ANN의 압축된 파라미터들을 저장하도록 더 구성되는, 신경 프로세싱 유닛.

청구항 34

제 31 항에 있어서,

상기 스케줄러는 상기 적어도 하나의 PE 및 상기 적어도 하나의 내부 메모리와 동작 가능하게 연결되고, 상기 제 1 또는 제 2 배치채널의 적어도 일부의 크기를 조정하도록 구성되는, 신경 프로세싱 유닛.

청구항 35

제 31 항에 있어서,

상기 적어도 하나의 PE 및 상기 적어도 하나의 내부 메모리 사이에 위치하는 활성화 함수 처리 유닛을 더 포함하고,

상기 활성화 함수 처리 유닛은 상기 제 1 배치채널 및 제 상기 2 배치채널에 대응하는 특징맵들을 순차적으로 처리하여 상기 제 1 배치채널 및 제 상기 2 배치채널에 대응하는 활성화맵들을 순차적으로 출력하도록 구성된, 신경 프로세싱 유닛.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망을 위한 방법 및 신경 프로세싱 유닛에 관한 것이다.

배 경 기 술

[0002] 인간은 인식(Recognition), 분류(Classification), 추론(Inference), 예측(Predict), 조작/의사결정(Control/Decision making) 등을 할 수 있는 지능을 갖추고 있다. 인공지능(artificial intelligence: AI)은 인간의 지능을 인공적으로 모방하는 것을 의미한다.

[0003] 인간의 뇌는 뉴런(Neuron)이라는 수많은 신경세포로 이루어져 있으며, 각각의 뉴런은 시냅스(Synapse)라고 불리는 연결부위를 통해 수백에서 수천 개의 다른 뉴런들과 연결되어 있다. 인간의 지능을 모방하기 위하여, 생물학적 뉴런의 동작원리와 뉴런 간의 연결 관계를 모델링한 것을, 인공신경망(Artificial Neural Network, ANN) 모델이라고 한다. 즉, 인공신경망은 뉴런들을 모방한 노드들을 레이어(Layer: 계층) 구조로 연결시킨, 시스템이다.

[0004] 이러한 인공신경망모델은 레이어 수에 따라 '단층 신경망'과 '다층 신경망'으로 구분한다.

[0005] 일반적인 다층신경망은 입력 레이어와 은닉 레이어, 출력 레이어로 구성되는데, (1) 입력 레이어(input layer)는 외부의 자료들을 받아들이는 레이어로서, 입력 레이어의 뉴런 수는 입력되는 변수의 수와 동일하다. (2) 은닉 레이어(hidden layer)는 입력 레이어와 출력 레이어 사이에 위치하며 입력 레이어로부터 신호를 받아 특성을 추출하여 출력층으로 전달한다. (3) 출력 레이어(output layer)는 은닉 레이어로부터 신호를 받아 외부로 출력한다. 뉴런 간의 입력신호는 0에서 1 사이의 값을 갖는 각각의 연결강도와 곱해진 후 합산된다. 합산 값이 뉴런의 임계치보다 크면 뉴런이 활성화되어 활성화 함수를 통하여 출력 값으로 구현된다.

[0006] 한편, 보다 높은 인공 지능을 구현하기 위하여, 인공신경망의 은닉 레이어의 개수를 늘린 것을 심층 신경망(Deep Neural Network, DNN)이라고 한다.

[0007] DNN에는 여러 종류가 있으나, 컨볼루션 신경망(Convolutional Neural Network, CNN)은 입력 데이터의 특징들을 추출하고, 특징들의 패턴을 파악하기에 용이한 것으로 알려져 있다.

[0008] 컨볼루션 신경망(CNN)은 인간 뇌의 시각 피질에서 영상을 처리하는 것과 유사한 기능을 하는 신경망이다. 컨볼루션 신경망은 영상처리에 적합한 것으로 알려져 있다.

[0009] 도 4를 참조하면, 컨볼루션 신경망은 컨볼루션 채널들과 풀링(pooling) 채널들이 반복되는 형태로 구성된다. 컨볼루션 신경망에서 대부분의 연산시간은 컨볼루션 동작이 차지한다. 컨볼루션 신경망은 행렬(Matrix) 형태의 커널(kernel)에 의해 각 채널의 영상의 특징을 추출하고, 풀링(Pooling)에 의해 이동이나 왜곡 등의 항상성을 제공하는 방식으로 사물을 인식한다. 각 채널에서는 입력 데이터와 커널의 컨볼루션으로 특징맵(Feature Map)을 구한 후 ReLU(Rectified Linear Unit) 같은 활성함수를 적용하여 해당 채널의 활성화 맵을 생성한다. 이후 풀링이 적용될 수 있다. 패턴을 실제로 분류하는 신경망은 특징 추출 신경망의 후단에 위치하며, 완전 연결 레이어(Fully Connected Layer)라고 한다. 컨볼루션 신경망의 연산 처리에서 대부분의 연산은 컨볼루션 또는 행렬곱을 통해 수행된다. 이때 필요한 커널들을 메모리로부터 읽어 오는 빈도가 상당히 빈번하다. 이러한 컨볼루션 신경망 동작의 상당 부분은 각각의 채널에 대응되는 커널들을 메모리로부터 읽어오는 시간이 차지한다.

[0010] 메모리는 메인 메모리, 내부 메모리, 온 칩(On-Chip) 메모리 등으로 나뉘어 진다. 각각의 메모리는 복수의 메모리 셀로 이루어지며, 각각의 메모리 셀은 고유한 메모리 주소를 가진다. 특히, 인공신경망 프로세서가 메인 메모리에 저장된 가중치 값을 불러오거나 다른 파라미터 값들을 불러올 때마다, 메인 메모리의 주소에 대응되는 메인 메모리 셀에 접근하기까지 여러 클럭(clock)의 지연시간(latency)이 발생될 수 있다.

[0011] 따라서 메인 메모리에서 필요한 파라미터를 읽어와 컨볼루션을 수행하는데 소모되는 시간과 전력 소모가 상당하다는 문제가 있다.

발명의 내용

해결하려는 과제

[0012] 본 개시의 발명자는 하기의 사항들에 대하여 인식하였다.

[0013] 먼저, 인공신경망모델의 추론 연산 시, 신경 프로세싱 유닛(NPU)이 빈번하게 인공신경망모델의 각각의 레이어의 노드 및/또는 가중치 값을 메인 메모리에서 읽어온다.

[0014] NPU는 인공신경망모델의 노드 및/또는 커널의 가중치 값 등을 메인 메모리에서 읽어오는 동작의 처리 속도가 느리고 에너지를 많이 소비한다.

[0015] 메인 메모리에 대한 액세스가 아닌 온칩 메모리나 NPU 내부 메모리에 대한 엑세스가 늘어날 수록 NPU의 처리 속도가 빨라지고 에너지 소비도 감소한다.

[0016] 복수의 채널을 하나의 NPU와 하나의 인공신경망모델로 처리하는 경우, 각각의 채널들을 개별적으로 처리할 때마다 동일한 가중치를 메인 메모리에서 반복적으로 읽어오는 것이 비효율 적이다.

[0017] 특히, 데이터가 일렬로 배치되어 처리되는 배치(batch) 채널들을 처리할 때, 그 처리 방식과 순서의 특징에 따라 온칩 메모리나 NPU 내부 메모리에 대한 활용을 극대화할 수 있다.

[0018] 마지막으로, 배치채널들의 컨볼루션 계산 처리에 있어서 반복 사용되는 파라미터들을 제한된 온칩 메모리나 NPU 내부 메모리에 최대한 유지시키는 것이 처리 속도를 극대화하고 에너지 소비도 감소시킬 수 있다.

[0019] 이에, 본 개시가 해결하고자 하는 과제는 온칩 메모리 또는 NPU 내부 메모리가 인공신경망의 파라미터들을 저장하고 계산하는 순서를 결정하여, 메인 메모리 읽기 동작의 횟수를 저감하고, 소비 전력을 저감할 수 있는 신경 프로세싱 유닛 및 그 동작 방법을 제공하는 것이다.

[0020] 또한, 본 개시가 해결하고자 하는 과제는 배치채널들의 처리가 빈번한 자율 자동차나 드론, 복수의 센서를 가지는 전자 디바이스 등에서 저전력으로 높은 성능을 가지는 신경 프로세싱 유닛 및 그 동작 방법을 제공하는 것이다.

[0021] 단 본 개시는 이에 제한되지 않으며, 또 다른 과제들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을

것이다.

과제의 해결 수단

[0022] 전술한 바와 같은 과제를 해결하기 위하여 본 발명의 일 예시에 따른 인공신경망을 위한 방법이 제공된다.

[0023] 상기 방법은 ANN (artificial neural network) 에 대한 동작들을 수행하는 단계를 포함하고, 동작들을 위해, 복수의 배치채널들은 제 1 배치채널 및 제 2 배치채널을 포함하고, 동작들은, 일 세트의 가중치, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부를 적어도 하나의 메모리에 저장하는 단계; 그리고 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치 값들을 계산하는 단계를 포함한다.

[0024] 제 1 배치채널의 적어도 일부의 크기 및 적어도 제 2 배치채널의 크기는 실질적으로 동일할 수 있다.

[0025] 상기 동작들을 위해, 일 세트의 가중치 값들은 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응할 수 있다.

[0026] 상기 동작들을 위해, 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 방법은, 제 2 배치채널의 적어도 일부와 일 세트 된 가중치 값들을 계산하는 동안, 적어도 하나의 메모리의 적어도 일부에 다음에 계산될 제 1 배치채널의 다른 적어도 일부를 저장하는 단계를 더 포함할 수 있다.

[0027] 상기 동작들을 위해, 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 더 포함하고, 동작들은, 일 세트의 가중치 값들을 유지하면서 적어도 하나의 메모리에 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부를 메모리에 저장하는 단계 및 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부와 일 세트의 가중치 값들을 계산하는 단계를 더 포함할 수 있다.

[0028] 상기 동작들을 위해, 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 그리고 일 세트의 가중치들은 대응되는 복수의 배치채널들 각각의 적어도 일부가 계산될 때까지 온-칩 메모리 및/또는 내부 메모리에 유지될 수 있다.

[0029] 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 동작들은 다음 세트의 가중치, 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분을 온-칩 메모리 및/또는 내부 메모리에 저장하는 단계 및 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분과 다음 세트의 가중치를 계산하는 단계를 포함할 수 있다.

[0030] 상기 동작들은 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치로부터 계산된 값들을 적어도 하나의 메모리에 저장하는 단계; 다음 프로세싱 단계를 위해 적어도 하나의 메모리에 다음 세트의 가중치들을 저장하는 단계; 그리고 계산된 값들과 다음 세트의 가중치를 계산하는 단계를 포함할 수 있다.

[0031] 적어도 하나의 메모리는 내부 메모리를 포함하고, 제 1 값들과 다음 세트의 가중치를 계산한 제 2 값들은 내부 메모리에서만 상주할 수 있다.

[0032] 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부는 완전한 데이터 세트를 포함할 수 있다.

[0033] 적어도 하나의 메모리는 내부 메모리를 포함하고, 동작들은, 내부 메모리에 저장하기 전에, 일 세트의 가중치 값들의 크기, 제 1 배치채널의 적어도 일부의 크기 및 제 2 배치채널의 적어도 일부의 크기 상기를 내부 메모리에 피팅(fitting)되도록 타일링(tiling)하는 단계를 포함할 수 있다.

[0034] 상기 동작들에 대해, ANN은 복수의 배치채널들로부터의 객체 검출, 분류 또는 세그먼트화를 포함하는 적어도 하나의 동작을 수행하도록 구성될 수 있다.

[0035] 상기 동작들을 위해, 객체는 차량, 신호등, 장애물, 통행인, 사람, 동물, 도로, 표지판, 및 도로 선 중 적어도 하나를 포함할 수 있다.

[0036] 상기 ANN에 대한 동작들 전에 복수의 배치채널들을 전처리하는 단계를 포함할 수 있다.

[0037] 상기 동작들을 위해, ANN은 객체들의 향상된 검출을 위해 복수의 배치채널들을 전처리하면서, 복수의 배치채널들로부터 객체들을 동시에 검출하도록 구성될 수 있다.

[0038] 상기 동작들에 대해, 복수의 배치채널들 각각은 복수의 이미지들 각각에 대응할 수 있다.

[0039] 상기 동작들을 위해, 복수의 배치채널들 중 적어도 하나의 배치채널은 IR, RGB, YCBCR, HSV, 및 HIS 형식일 수

있다.

[0040] 상기 동작들을 위해, 복수의 배치채널들 중 적어도 하나는 차량의 내부를 캡처하는 이미지를 포함하고, ANN은 차량의 안전 관련 객체, 기능, 운전자의 상태, 및 승객의 상태 중 적어도 하나를 검출하도록 구성될 수 있다.

[0041] 상기 동작들을 위해, 복수의 이미지들은 RGB 이미지, IR 이미지, RADAR 이미지, ULTRASOUND 이미지, LIDAR 이미지, 열 화상 이미지, NIR 이미지 및 이들의 융합 이미지 중 적어도 하나를 포함할 수 있다.

[0042] 상기 동작들을 위해, 복수의 이미지들은 실질적으로 동일한 기간에 캡처된 이미지들이다.

[0043] 상기 동작들을 위해, 복수의 배치채널들 각각은 복수의 센서 데이터 각각에 대응하고, 그리고 복수의 센서 데이터는 압력 센서, 피에조 센서, 습도 센서, 먼지 센서, 스모그 센서, 소나 센서, 진동 센서, 가속도 센서 또는 모션 센서 중 하나 이상으로부터의 데이터를 포함할 수 있다.

[0044] 전술한 바와 같은 과제를 해결하기 위하여 본 발명의 다른 예시에 따른 신경 프로세싱 유닛이 제공된다. 신경 프로세싱 유닛은 상기 제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크에 대한 신경 프로세싱 유닛으로서, 제 1 배치채널의 적어도 일부, 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치 값들을 저장하도록 구성된 적어도 하나의 내부 메모리; 그리고 저장된 일 세트의 가중치 값들을 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element)를 포함한다.

[0045] 적어도 하나의 내부 메모리에 할당되는 제 1 배치채널의 적어도 일부의 크기 및 적어도 제 2 배치채널의 크기는 실질적으로 동일할 수 있다.

[0046] 일 세트의 가중치는 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응할 수 있다.

[0047] 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 포함하고, 적어도 하나의 내부 메모리는, 일 세트의 가중치를 유지하는 동안, 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부를 저장하도록 더 구성되고, PE는 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부 각각과 상기 일 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0048] 적어도 하나의 내부 메모리는 복수의 배치채널들이 계산될 때까지 일 세트의 가중치를 유지하도록 더 구성될 수 있다.

[0049] 적어도 하나의 내부 메모리는 또 다른 세트의 가중치, 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분을 저장하도록 더 구성되고; 그리고 PE는 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분 각각과 또 다른 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0050] 적어도 하나의 내부 메모리는 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치로부터 계산된 값들을 저장하고, 그리고 다음 스테이지에 대한 일 세트의 가중치를 저장하도록 더 구성되고, PE는 계산된 값들과 다음 스테이지에 대한 일 세트의 가중치를 계산하도록 더 구성되고, 일 세트의 가중치는 복수의 배치채널들이 계산될 때까지 내부 메모리에 유지하도록 더 구성될 수 있다.

[0051] 적어도 하나의 내부 메모리는, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응하고, 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치 로부터 계산된 제 1 계산 값들을 저장하고, 다음 프로세싱 단계를 위한 다음 세트의 가중치를 저장하도록 더 구성되고, PE는 제 1 계산 값들과 다음 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0052] 신경 프로세싱 유닛은 일 세트의 가중치의 크기, 제 1 배치채널의 적어도 일부의 크기 및 제 2 배치채널의 적어도 일부의 크기를 내부 메모리에 맞게 조정하도록 구성된 스케줄러를 더 포함할 수 있다.

[0053] 전술한 바와 같은 과제를 해결하기 위하여 본 발명의 또 다른 예시에 따른 신경 프로세싱 유닛이 제공된다. 신경 프로세싱 유닛은 제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크 (ANN) 를 위한 신경 프로세싱 유닛으로서, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치 값들을 저장하도록 구성된 적어도 하나의 내부 메모리 및 저장된 일 세트의 가중치 값들을 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element) 를 포함하고, 제 1 배치채널의 적어도 일부의 크기는, 적어도 하나의 내부 메모리의 크기를 복수의 채널들의 수로 나눈 것과 같거나 작을 수 있다.

[0054] 적어도 하나의 내부 메모리의 크기는 ANN의 가장 큰 특징맵의 크기 및 배치채널들의 수에 대응할 수 있다.

[0055] 적어도 하나의 내부 메모리는 ANN의 압축된 파라미터들을 저장하도록 더 구성될 수 있다.

[0056] 적어도 하나의 PE 및 적어도 하나의 내부 메모리와 동작 가능하게 연결되고, 제 1 또는 제 2 배치채널의 적어도 일부의 크기를 조정하도록 구성된 스케줄러를 더 포함할 수 있다.

[0057] 신경 프로세싱 유닛은 적어도 하나의 PE 및 적어도 하나의 내부 메모리 사이에 위치하는 활성화 함수 처리 유닛을 더 포함하고, 활성화 함수 처리 유닛은 제 1 배치채널 및 제 2 배치채널에 대응하는 특징맵들을 순차적으로 처리하여 제 1 배치채널 및 제 2 배치채널에 대응하는 활성화맵들을 순차적으로 출력할 수 있다.

발명의 효과

[0058] 본 개시에 따르면, 복수의 입력 채널을 처리하도록 구성된 배치모드로 온칩 메모리 및/또는 내부 메모리가 인공신경망의 파라미터들을 저장하고 계산하는 순서를 결정함으로써, 메인 메모리 읽기 동작의 횟수를 저감하고, 소비 전력을 저감할 수 있다.

[0059] 본 개시에 따르면, 입력 채널 개수가 증하더라도, 복수의 입력 채널을 고려하여 구성된 온칩 메모리 및/또는 내부 메모리를 포함하는 하나의 신경 프로세싱 유닛으로 처리할 수 있다.

[0060] 또한, 본 개시에 따르면, 배치채널들의 처리가 빈번한 자율 자동차나 드론, 복수의 센서를 가지는 전자 디바이스 등에서 저전력으로 높은 성능을 가지는 신경 프로세싱 유닛을 제공할 수 있다.

[0061] 또한, 본 개시에 따르면, 배치채널들의 개수 및 연산 성능을 고려하여 온칩 메모리 또는 내부 메모리의 크기를 결정한 배치모드 전용 신경 프로세싱 유닛을 제공할 수 있다.

도면의 간단한 설명

[0062] 도 1은 본 개시의 일 예시에 따른 신경 프로세싱 유닛이 포함된 장치를 설명하는 개략적인 개념도이다.

도 2a는 본 개시의 일 예시에 따른 신경 프로세싱 유닛을 설명하는 개략적인 개념도이다.

도 2b는 NPU(1000)의 동작 시 소모되는 에너지를 나타낸 예시도이다.

도 2c는 본 개시에 적용될 수 있는 프로세싱 엘리먼트 어레이 중 하나의 프로세싱 엘리먼트를 설명하는 개략적인 개념도이다.

도 3은 도 2a에 도시된 NPU(1000)의 변형예를 나타낸 예시도이다.

도 4는 예시적인 인공신경망모델을 설명하는 개략적인 개념도이다.

도 5는 본 개시의 일 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다.

도 6은 본 개시의 일 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

도 7은 본 개시의 다른 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다.

도 8은 본 개시의 다른 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

도 9는 본 개시의 또 다른 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다.

도 10은 본 개시의 또 다른 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

도 11은 본 개시의 다양한 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다.

도 12는 본 개시의 다양한 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

도 13은 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템을 나타낸 예시도이다.

도 14는 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템의 개략적인 블록도이다.

도 15는 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템에서 자율 주행을 위해 목표 객체를 인식하기 위한 발명을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0063] 본 명세서 또는 출원에 개시되어 있는 본 개시의 개념에 따른 실시 예들에 대해서 특정한 구조적 내지 단계적 설명들은 단지 본 개시의 개념에 따른 실시 예를 설명하기 위한 목적으로 예시된 것이다.

[0064] 본 개시의 개념에 따른 실시 예들은 다양한 형태로 실시될 수 있으며 본 개시의 개념에 따른 실시 예들은 다양한 형태로 실시될 수 있으며 본 명세서 또는 출원에 설명된 실시 예들에 한정되는 것으로 해석되어서는 아니 된다.

[0065] 본 개시의 개념에 따른 실시 예는 다양한 변경을 가할 수 있고 여러 가지 형태를 가질 수 있으므로 특정 실시 예들을 도면에 예시하고 본 명세서 또는 출원에 상세하게 설명하고자 한다. 그러나, 이는 본 개시의 개념에 따른 실시 예를 특정한 개시 형태에 대해 한정하려는 것이 아니며, 본 개시의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0066] 제 1 및/또는 제 2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다.

[0067] 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 개시의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소는 제 1 구성요소로도 명명될 수 있다.

[0068] 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.

[0069] 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0070] 본 문서에서, "A 또는 B," "A 또는/및 B 중 적어도 하나," 또는 "A 또는/및 B 중 하나 또는 그 이상" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다. 예를 들면, "A 또는 B," "A 및 B 중 적어도 하나," 또는 "A 또는 B 중 적어도 하나"는, (1) 적어도 하나의 A를 포함, (2) 적어도 하나의 B를 포함, 또는(3) 적어도 하나의 A 및 적어도 하나의 B 모두를 포함하는 경우를 모두 지칭할 수 있다.

[0071] 본 문서에서 사용된 "제 1," "제 2," "첫째," 또는 "둘째," 등의 표현들은 다양한 구성요소들을, 순서 및/또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다. 예를 들면, 제 1 사용자 기기와 제 2 사용자 기기는, 순서 또는 중요도와 무관하게, 서로 다른 사용자 기기를 나타낼 수 있다. 예를 들면, 본 문서에 기재된 권리범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 바꾸어 명명될 수 있다.

[0072] 본 문서에서 사용된 용어들은 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 다른 예시의 범위를 한정하려는 의도가 아닐 수 있다.

[0073] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다. 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 용어들은 본 문서에 기재된 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가질 수 있다.

[0074] 본 문서에 사용된 용어들 중 일반적인 사전에 정의된 용어들은, 관련 기술의 문맥상 가지는 의미와 동일 또는 유사한 의미로 해석될 수 있으며, 본 문서에서 명백하게 정의되지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다. 경우에 따라서, 본 문서에서 정의된 용어일지라도 본 문서의 실시 예들을 배제하도록 해석될 수 없다.

[0075] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 개시를 한정하려는 의도가 아니다.

[0076] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다"

또는 "가지다" 등의 용어는 서술된 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.

[0077] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0078] 본 발명의 여러 예시들의 각각 특징들이 부분적으로 또는 전체적으로 서로 결합 또는 조합 가능하며, 당업자가 충분히 이해할 수 있듯이 기술적으로 다양한 연동 및 구동이 가능하며, 각 예시들이 서로에 대하여 독립적으로 실시 가능할 수도 있고 연관 관계로 함께 실시 가능할 수도 있다.

[0079] 실시 예를 설명함에 있어서 본 개시가 속하는 기술 분야에 익히 알려져 있고 본 개시와 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략한다. 이는 불필요한 설명을 생략함으로써 본 개시의 요지를 흐리지 않고 더욱 명확히 전달하기 위함이다.

[0080] <용어의 정의>

[0081] 이하, 본 명세서에서 제시되는 개시들의 이해를 돕고자, 본 명세서에서 사용되는 용어들에 대하여 간략하게 정리하기로 한다.

[0082] NPU: 신경 프로세싱 유닛(Neural Processing Unit)의 약어로서, CPU(Central processing unit)과 별개로 인공신경망모델의 연산을 위해 특화된 프로세서를 의미할 수 있다. 인공신경망 가속기로 지칭되는 것도 가능하다.

[0083] NPU 스케줄러(또는 스케쥴러): NPU 스케줄러는 NPU의 전반적인 타스크(task)를 제어하는 모듈을 의미할 수 있다. NPU 스케줄러는 NPU에서 구동을 위해, 컴파일러가 ANN모델의 데이터 지역성을 분석하여 컴파일된 ANN 모델의 연산순서 정보를 제공받아 NPU의 업무 처리 순서를 결정한다. NPU 스케줄러는 정적인 ANN 모델의 데이터 지역성을 기초로 결정된 정적인 타스크 순서로 NPU를 제어할 수 있다. NPU 스케줄러는 동적으로 ANN 모델의 데이터 지역성을 분석하여 동적인 타스크 순서로 NPU를 제어할 수 있다. NPU 스케줄러에는 NPU의 메모리 크기 및 프로세싱 엘리먼트 어레이의 성능을 기초로 ANN 모델의 레이어별 타일링 정보가 저장될 수 있다. NPU 스케줄러는 레지스터맵을 이용하여 NPU의 전반적인 타스크를 제어할 수 있다. NPU 스케줄러는 NPU에 포함되거나, NPU 외부에 배치될 수 있다.

[0084] ANN: 인공신경망(artificial neural network)의 약어로서, 인간의 지능을 모방하기 위하여, 인간 뇌 속의 뉴런들(Neurons)이 시냅스(Synapse)를 통하여 연결되는 것을 모방하여, 노드들을 레이어(Layer: 계층) 구조로 연결시킨, 네트워크를 의미할 수 있다.

[0085] 인공신경망의 구조에 대한 정보: 레이어의 개수에 대한 정보, 레이어 내의 노드의 개수, 각 노드의 값, 연산 처리 방법에 대한 정보, 각 노드에 적용되는 가중치 행렬에 대한 정보 등을 포함하는 정보이다.

[0086] 인공신경망모델의 데이터 지역성: 학습이 완료된 인공신경망(ANN)의 구조가 확정되면, 인공신경망모델을 컴파일하여 확정된 모든 연산순서 및 연산 종류를 포함하는 정보이다.

[0087] DNN: 심층 신경망(Deep Neural Network)의 약어로서, 보다 높은 인공 지능을 구현하기 위하여, 인공신경망의 은닉 레이어의 개수를 늘린 것을 의미할 수 있다.

[0088] CNN: 컨볼루션 신경망(Convolutional Neural Network)의 약어로서, 인간 뇌의 시각 피질에서 영상을 처리하는 것과 유사한 기능을 하는 신경망이다. 컨볼루션 신경망은 영상처리에 적합한 것으로 알려져 있으며, 입력 데이터의 특징들을 추출하고, 특징들의 패턴을 파악하기에 용이한 것으로 알려져 있다. CNN에서의 가중치는 N x M 크기의 커널을 지칭할 수 있다.

[0089] Fused-ANN: 융합 신경망(Fused Artificial Neural Network)의 약어로서, 센서 융합된 데이터를 처리하도록 설계된 인공신경망을 의미할 수 있다. 센서 융합은 자율 주행 기술 분야에서 주로 활용된다. 센서 융합은 하나의 센서가 특정 조건에서 감지 성능이 낮아질 경우, 다른 종류의 센서가 보완해 주는 기술일 수 있다. 센서 융합은 카메라와 열화상 카메라의 융합, 카메라와 레이더의 융합, 카메라와 라이더의 융합, 카메라와 레이더와 라이더의 융합 등, 그 경우의 수가 다양할 수 있다. 융합 신경망은 SKIP-CONNECTION, SQUEEZE-AND-EXCITATION,

CONCATENATION 등의 오퍼레이터를 더 추가하여 다수의 센서 데이터를 융합시킨 인공신경망모델일 수 있다.

[0090] 이하, 첨부한 도면을 참조하여 본 개시의 실시 예를 설명함으로써, 본 개시를 상세히 설명한다. 이하, 본 개시의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0091] 도 1은 본 개시의 일 예시에 따른 신경 프로세싱 유닛이 포함된 장치를 설명하는 개략적인 개념도이다.

[0092] 도 1을 참조하면 NPU(1000)가 포함된 장치(B)는 온칩 영역(A)을 포함한다. 온칩 영역 외부에는 메인 메모리(4000)가 포함될 수 있다. 메인 메모리(4000)는 예를 들어 DRAM등과 같은 시스템 메모리일 수 있다. 도시되지 않았으나, 온칩 영역(A) 외부에는 ROM등을 포함하는 저장부가 포함될 수 있다.

[0093] 온칩 영역(A)에는 중앙 프로세싱 유닛(CPU)(2000)과 같은 범용 프로세싱 유닛과 온칩 메모리(3000) 그리고 NPU(1000)가 배치된다. CPU(2000)는 NPU(1000)와 온칩 메모리(3000) 그리고 메인 메모리(4000)에 동작 가능하게 연결된다.

[0094] 단, 본 개시는 이에 제한되지 않으며, CPU(2000) 내부에 NPU(1000)가 포함되도록 구성되는 것도 가능하다.

[0095] 온칩 메모리(3000)는 반도체 다이에 실장된 메모리로 메인 메모리(4000) 액세스와 별도로 캐싱을 위한 메모리일 수 있다.

[0096] 예를 들면, 온칩 메모리(3000)는 다른 온칩 반도체들이 액세스하도록 설정된 메모리일 수도 있다. 예를 들면, 온칩 메모리(3000)는 캐시 메모리 또는 버퍼 메모리 일 수 있다.

[0097] NPU(1000)는 내부 메모리(200)를 포함하며, 내부 메모리(200)는 예를 들어 SRAM을 포함할 수 있다. 내부 메모리(200)는 실질적으로 NPU(1000)에서의 연산에만 사용되는 메모리 일 수 있다. 내부 메모리(200)는 NPU 내부 메모리로 지칭될 수 있다. 여기서 실질적이란, 내부 메모리(200)에는 NPU(1000)가 처리하는 인공신경망과 관련된 데이터를 저장하도록 구성된 것을 의미할 수 있다.

[0098] 예를 들면, 내부 메모리(200)는 NPU(1000) 연산에 필요한 가중치, 커널 및/또는 특징맵을 저장하도록 구성된 버퍼 메모리 및/또는 캐시 메모리 일 수 있다. 단, 이에 제한되지 않는다.

[0099] 예를 들면, 내부 메모리(200)는 SRAM, MRAM, 레지스터 파일(Register file) 등의 읽고 쓰기가 메인 메모리(4000)보다 상대적으로 더 빠른 메모리 소자로 구성될 수 있다. 단, 이에 제한되지 않는다.

[0100] NPU(1000)가 포함된 장치(B)는 내부 메모리(200), 온칩 메모리(3000), 메인 메모리(4000) 중 적어도 하나를 포함한다.

[0101] 이하에서 설명하는 “적어도 하나의 메모리”는 내부 메모리(200), 및 온칩 메모리(3000)중 적어도 하나를 포함하도록 의도된다.

[0102] 또한, 온칩 메모리(3000)의 기재는 NPU(1000)의 내부 메모리(200) 또는 NPU(1000)의 외부에 있으나 온칩 영역(A)에 있는 메모리를 포함하도록 의도될 수 있다.

[0103] 다만, 적어도 하나의 메모리를 지칭하는 내부 메모리(200) 및/또는 온칩 메모리(3000)는 위치적 특성이 아닌 메모리의 대역폭(bandwidth) 기준으로 메인 메모리(4000)와 구분하는 것도 가능하다.

[0104] 통상적으로 메인 메모리(4000)는 대용량의 데이터를 저장하기 용이하나, 메모리 대역폭이 상대적으로 낮고, 전력 소모가 상대적으로 큰 메모리를 지칭한다.

[0105] 통상적으로 내부 메모리(200)와 온칩 메모리(3000)는 메모리 대역폭이 상대적으로 높고, 전력 소모가 상대적으로 낮으나, 대용량의 데이터를 저장하기에 비효율적인 메모리를 지칭한다.

[0106] NPU(1000)가 포함된 장치(B)의 각각의 구성요소는 버스(5000)를 통해서 통신할 수 있다. 장치(B)의 버스(5000)는 적어도 하나일 수 있다. 버스(5000)는 통신 버스, 및/또는 시스템 버스 등으로 지칭될 수 있다.

[0107] NPU(1000)의 내부 메모리(200)와 온 칩 메모리(3000)는 인공신경망모델의 가중치와 특징맵 처리를 위해 특정 대역폭 이상을 보장하기 위해서 별도의 전용 버스를 더 구비하는 것도 가능하다.

[0108] 온 칩 메모리(3000)와 메인 메모리(4000) 사이에는 특정 대역폭 이상을 보장하기 위해서 별도의 전용 버스를 더 구비하는 것도 가능하다. 상기 특정 대역폭은 NPU(1000)의 프로세싱 엘리먼트 어레이의 처리 성능을 기준으로 결정될 수 있다.

[0109] NPU(1000)의 내부 메모리(200)와 메인 메모리(4000) 사이에는 특정 대역폭 이상을 보장하기 위해서 별도의 전용 버스를 더 구비하는 것도 가능하다. 상기 특정 대역폭은 NPU(1000)의 프로세싱 엘리먼트 어레이의 처리 성능을 기준으로 결정될 수 있다.

[0110] NPU(1000)가 포함된 장치(B)는 DMA(Direct Memory Access) 모듈을 더 포함하여, 내부 메모리(200), 온 칩 메모리(3000) 및/또는 메인 메모리(4000)를 직접 제어하도록 구성되는 것도 가능하다.

[0111] 예를 들면, DMA 모듈은 버스(5000)를 직접 제어하여 NPU(1000)와 온칩 메모리(3000)의 데이터 전송을 직접 제어하도록 구성될 수 있다.

[0112] 예를 들면, DMA 모듈은 버스(5000)를 직접 제어하여 온칩 메모리(3000)와 메인 메모리(4000)의 데이터 전송을 직접 제어하도록 구성될 수 있다.

[0113] 예를 들면, DMA 모듈은 버스(5000)를 직접 제어하여 내부 메모리(200)와 메인 메모리(4000)의 데이터 전송을 직접 제어하도록 구성될 수 있다.

[0114] 신경 프로세싱 유닛(neural processing unit, NPU)(1000)은 인공신경망을 위한 동작을 수행하도록 특화된 프로세서이다. NPU(1000)는 AI 가속기로 지칭될 수 있다.

[0115] 인공신경망은 여러 입력 또는 자극이 들어오면 각각 가중치를 곱해 더해주고, 추가적으로 편차를 더한 값을 활성화 함수를 통해 변형하여 전달하는 인공 뉴런들이 모인 네트워크를 의미한다. 이렇게 학습된 인공신경망은 입력 데이터로부터 추론(inference) 결과를 출력하는데 사용될 수 있다.

[0116] 상기 NPU(1000)는 전기/전자 회로로 구현된 반도체일 수 있다. 상기 전기/전자 회로라 함은 수많은 전자 소자, (예컨대 트렌지스터, 커패시터)를 포함하는 것을 의미할 수 있다. 상기 NPU(1000)는 프로세싱 엘리먼트(processing element: PE) 어레이, NPU 내부 메모리(200), NPU 스케줄러, 및 NPU 인터페이스를 포함할 수 있다. 프로세싱 엘리먼트 어레이, NPU 내부 메모리(200), NPU 스케줄러, 및 NPU 인터페이스 각각은 수많은 트렌지스터들이 연결된 반도체 회로일 수 있다.

[0117] 따라서, 이들 중 일부는 육안으로는 식별되어 구분되기 어려울 수 있고, 동작에 의해서만 식별될 수 있다. 예컨대, 임의 회로는 프로세싱 엘리먼트 어레이로 동작하기도 하고, 혹은 NPU 스케줄러로 동작될 수도 있다.

[0118] 상기 NPU(1000)는 프로세싱 엘리먼트 어레이, 프로세싱 엘리먼트 어레이에서 추론될 수 있는 인공신경망모델의 적어도 일부를 저장하도록 구성된 NPU 내부 메모리(200), 및 인공신경망모델의 데이터 지역성 정보 또는 인공신경망모델의 구조에 대한 정보에 기초하여 프로세싱 엘리먼트 어레이 및 NPU 내부 메모리(200)를 제어하도록 구성된 NPU 스케줄러를 포함할 수 있다.

[0119] 인공신경망모델은 인공신경망모델의 데이터 지역성 정보 또는 구조에 대한 정보를 포함할 수 있다.

[0120] 인공신경망모델은 특정 추론 기능을 수행하도록 학습된 AI 인식모델을 의미할 수 있다.

[0121] 프로세싱 엘리먼트 어레이는 인공신경망을 위한 동작을 수행할 수 있다. 예를 들어, 입력 데이터가 입력되었을 때, 프로세싱 엘리먼트 어레이는 인공신경망이 학습을 수행하도록 할 수 있다. 학습이 완료된 이후, 입력 데이터가 입력되었을 때, 프로세싱 엘리먼트 어레이는 학습 완료된 인공신경망을 통해 추론 결과를 도출하는 동작을 수행할 수 있다.

[0122] 예를 들면, NPU(1000)는 NPU 인터페이스를 통해서 메인 메모리(4000)에 저장된 인공신경망모델의 데이터를 NPU 내부 메모리(200)로 불러올 수 있다. NPU 인터페이스는 버스(5000)를 통해서 메인 메모리(4000)와 통신할 수 있다.

[0123] NPU 스케줄러는 NPU(1000)의 추론 연산을 위한 프로세싱 엘리먼트 어레이의 연산 및 NPU 내부 메모리(200)의 읽기 및 쓰기 순서를 제어하도록 구성된다. 또한 NPU 스케줄러는 배치채널의 적어도 일부의 크기를 조정하도록 구성된다.

[0124] NPU 스케줄러는 인공신경망모델의 구조를 분석하거나 또는 인공신경망모델의 구조를 제공받는다. 다음으로, NPU 스케줄러는 각 레이어 별 연산순서를 순차적으로 결정한다. 즉, 인공신경망모델의 구조가 확정될 경우, 레이어 별 연산순서가 정해질 수 있다. 이러한 인공신경망모델의 구조에 따른 연산의 순서 또는 데이터 흐름의 순서를 알고리즘 레벨에서의 인공신경망모델의 데이터 지역성으로 정의할 수 있다.

[0125] NPU 스케줄러는 상기 인공신경망모델의 구조와 배치채널의 개수를 반영하여 각 레이어 별 연산순서를 순차적으

로 결정한다. 즉, 인공신경망모델의 구조 및 배치채널의 개수가 확정될 경우, 레이어 별 연산순서가 정해질 수 있다. 이러한 배치채널의 개수 및 인공신경망모델의 구조에 따른 연산의 순서 또는 데이터 흐름의 순서를 알고리즘 레벨에서의 인공신경망모델의 데이터 지역성 또는 배치모드의 인공신경망모델의 데이터 지역성으로 정의할 수 있다. 이하 배치모드의 인공신경망 모델의 데이터 지역성은 인공신경망 모델의 데이터 지역성으로 지칭될 수 있다.

[0126] 인공신경망모델의 데이터 지역성은 인공신경망모델의 구조, 배치채널의 개수, 및 NPU 구조를 모두 고려하여 결정될 수 있다.

[0127] 인공신경망모델이 NPU(1000)에서 실행되도록 컴파일러가 인공신경망모델을 컴파일할 경우, 신경 프로세싱 유닛-메모리 레벨에서의 인공신경망모델의 인공신경망 데이터 지역성이 재구성될 수 있다. 예를 들어, 컴파일러는 CPU(2000)에 의해 실행될 수 있다.

[0128] 즉, 컴파일러, 인공신경망모델에 적용된 알고리즘들, 및 NPU(1000)의 동작 특성, 가중치 값들의 크기, 및 특징맵 또는 배치채널의 크기에 따라서 내부 메모리에 로딩되는 가중치 값들, 및 배치채널의 크기가 결정될 수 있다.

[0129] 예를 들면, 동일한 인공신경망모델의 경우에도 NPU(1000)가 해당 인공신경망모델을 연산하는 방식, 예를 들면, 특징맵 타일링(feature map tiling), 프로세싱 엘리먼트의 스테이셔너리(Stationary) 기법 등, NPU(1000)의 프로세싱 엘리먼트 개수, NPU(1000) 내 특징맵 및 가중치의 크기, 내부 메모리 용량, NPU(1000)내의 메모리 계층 구조, 및 해당 인공신경망모델을 연산 처리하기 위한 NPU(1000)의 연산 동작의 순서를 결정해 주는 컴파일러의 알고리즘 특성 등에 따라서 처리하고자 하는 인공신경망모델의 계산 방법이 구성될 수 있다. 왜냐하면, 상술한 요인들에 의해서 동일한 인공신경망모델을 연산 처리하더라도 NPU(1000)가 클럭 단위로 매 순간 필요한 데이터의 순서를 상이하게 결정할 수 있기 때문이다.

[0130] 도 2a는 본 개시의 일 예시에 따른 신경 프로세싱 유닛을 설명하는 개략적인 개념도이다.

[0131] 신경 프로세싱 유닛(NPU)(1000)은 스케줄러(300), 프로세싱 엘리먼트 어레이(100), 및 내부 메모리(200)를 포함한다.

[0132] NPU 스케줄러(300)는 인공신경망모델의 가중치 값들의 크기, 특징맵의 크기, 및 가중치 값들과 특징맵의 계산 순서 등을 고려하여 프로세싱 엘리먼트 어레이(100) 및 NPU 내부 메모리(200)를 제어하도록 구성될 수 있다.

[0133] NPU 스케줄러(300)는 프로세싱 엘리먼트 어레이(100)에서 계산될 가중치 값들의 크기, 특징맵의 크기, 및 가중치 값들과 특징맵의 계산 순서 등을 수신할 수 있다. 인공신경망모델이 포함할 수 있는 인공신경망의 데이터는 각각의 레이어의 노드 데이터 또는 특징맵, 및 각각의 레이어의 노드를 연결하는 연결망 각각의 가중치 데이터를 포함할 수 있다. 인공신경망의 데이터 또는 파라미터들 중 적어도 일부는 NPU 스케줄러(300) 내부에 제공되는 메모리 또는 NPU 내부 메모리(200)에 저장될 수 있다.

[0134] 인공신경망의 파라미터들 중 특징맵은 배치채널로 구성될 수 있다. 여기서 복수의 배치채널들은 예를 들어 실질적으로 동일한 기간, (예를 들어 10 또는 100 ms 이내)에 복수의 이미지 센서를 통해 촬영된 이미지들일 수 있다.

[0135] NPU 스케줄러(300)는 예를 들어 인공신경망의 컨볼루션 연산을 위해 프로세싱 엘리먼트 어레이(100)와 내부 메모리(200)를 제어할 수 있다. 먼저 NPU 스케줄러(300)는 일 세트의 가중치 값들을 내부 메모리(200)의 가중치 저장부(210)에 로드(load)하고, 상기 일 세트의 가중치 값들에 대응하는 복수의 배치채널들의 일부를 내부 메모리(200)의 배치채널 저장부(220)에 로드할 수 있다. NPU 스케줄러(300)는 일 세트의 가중치 값들과 복수의 배치채널들의 일부가 계산된 후 일 세트의 가중치 값들을 내부 메모리(200)에 유지하면서 다음 계산될 복수의 배치채널들을 로드할 수 있다. 내부 메모리(200)가 가중치 저장부(210) 및 배치채널 저장부(220)를 구분하여 포함하는 것으로 도시되었으나, 이는 예시적일 뿐이고, 메모리 주소 등을 통해 논리적으로 구분되거나 또는 가변적으로 구분되거나 또는 구분되지 않을 수도 있다.

[0136] 다양한 예시에서, 일 세트의 가중치 값들은 전체 가중치 값들의 일부 일 수 있다. 이러한 경우 복수의 배치채널들의 일부 예컨대 제 1 배치채널의 일부 및 제 2 배치채널의 일부가 먼저 계산되고, 제 1 배치채널의 다음 일부 및 제 2 배치채널의 다음 일부가 다음으로 계산될 수도 있다. 또는 복수의 배치채널들의 일부 예컨대 제 1 배치채널의 일부 및 제 2 배치채널의 일부가 먼저 계산되고, 제 3 배치채널의 일부 및 제 4 배치채널의 일부가 다음으로 계산될 수도 있다.

[0137] 다양한 예시에서, 일 세트의 가중치 값들이 제 2 배치채널의 일부와 계산되는 동안, 이미 계산된 제 1 배치채널의 일부 자리에 다음 계산될 제 3 배치채널의 일부가 로딩될 수도 있다. 계산과 동시에 다음 계산될 파라미터가 내부 메모리에 로딩되는 경우 처리 속도는 더 빨라 질 수 있다.

[0138] 위에 설명된 예시에서는 인공신경망의 파라미터들이 NPU의 내부 메모리(200)에 저장되는 것으로 설명되었지만, 이에 제한되지 않고 온칩 메모리 또는 메인 메모리에 저장될 수도 있다.

[0139] 본 개시의 NPU(1000)에서 처리 속도를 향상시키는 구성은 특히 메모리(어떤 종류의 메모리이더라도)에 가중치 값들을 저장한 후 최대한 추가적인 메모리 액세스 없이 유지시킴으로써, 도 2b를 참조하여 설명할 DRAM 메모리 또는 메인 메모리 읽기를 최소화하는 것이다. 가중치 값들 또는 특징맵들에 대한 메인 메모리 읽기 횟수는 에너지 소비와 비례하고, 처리 속도에 반비례하므로, 이들에 대한 메인 메모리 읽기 횟수를 줄이면 에너지 소비를 줄이면서 처리속도를 높일 수 있다.

[0140] 일반적인 CPU의 스케줄링은 공평성, 효율성, 안정성, 반응 시간 등을 고려하여, 최상의 효율을 낼 수 있도록 동작한다. 즉, 우선 순위, 연산 시간 등을 고려해서 동일 시간내에 가장 많은 프로세싱을 수행하도록 스케줄링 된다.

[0141] 종래의 CPU는 각 프로세싱의 우선 순서, 연산 처리 시간 등의 데이터를 고려하여 작업을 스케줄링 하는 알고리즘을 사용하였다.

[0142] 이와 다르게 NPU 스케줄러(300)는 인공신경망모델의 파라미터들의 계산 방식 특히, 배치채널들과 가중치들 사이의 계산의 특성에 기초하여 프로세싱 순서를 결정할 수 있다.

[0143] 더 나아가면, NPU 스케줄러(300)는 하나의 컨볼루션 연산이 끝날 때까지 하나의 가중치 세트가 모든 배치채널에 대해서 적용되어야 한다는 것에 근거하여, 가중치 세트를 메인 메모리에서 다시 액세스하지 않도록 프로세싱 순서를 결정할 수 있다.

[0144] 다르게 설명하면, 배치모드에서 하나의 컨볼루션 연산은 일 세트의 가중치로 순차적 복수의 배치채널들을 각각 컨볼루션하는 것을 의미할 수 있다.

[0145] 단, 본 개시는 NPU(1000)의 위의 근거에 제한되지 않고, 데이터 지역성 정보 또는 구조에 대한 정보에 더 기초할 수 있다. 예를 들면, NPU(1000)의 데이터 지역성 정보 또는 구조에 대한 정보는 NPU 내부 메모리(200)의 메모리 크기, NPU 내부 메모리(200)의 계층(hierarchy) 구조, 프로세싱 엘리먼트들(PE1 to PE12)의 개수 데이터, 프로세싱 엘리먼트들(PE1 to PE12)의 연산기 구조 중 적어도 하나 이상의 데이터를 포함할 수 있다. NPU 내부 메모리(200)의 메모리 크기는 메모리 용량에 대한 정보를 포함한다. NPU 내부 메모리(200)의 계층(hierarchy) 구조는 각각의 계층 구조에 대한 구체적인 계층 간의 연결 관계에 대한 정보를 포함한다. 프로세싱 엘리먼트들(PE1 to PE12)의 연산기 구조는 프로세싱 엘리먼트 내부의 구성요소들에 대한 정보를 포함한다.

[0146] 즉, NPU 스케줄러(300)는 NPU 내부 메모리(200)의 메모리 크기, NPU 내부 메모리(200)의 계층(hierarchy) 구조, 프로세싱 엘리먼트들(PE1 to PE12)의 개수 데이터, 프로세싱 엘리먼트들(PE1 to PE12)의 연산기 구조 중 적어도 하나 이상의 데이터를 활용하여 프로세싱 순서를 결정할 수 있다.

[0147] 단, 본 개시는 NPU(1000)에 제공되는 데이터 지역성 정보 또는 구조에 대한 정보에 제한되지 않는다.

[0148] 본 개시의 일 예시에 따른, NPU 스케줄러(300)는 인공신경망모델의 파라미터들의 계산 방식 특히, 배치채널들과 가중치들 사이의 계산의 특성에 기초하여 적어도 하나의 프로세싱 엘리먼트 및 NPU 내부 메모리(200)를 제어할 수 있다

[0149] 한편, 프로세싱 엘리먼트 어레이(100)는 인공신경망의 노드 데이터(예를 들면, 특징맵)와 연결망의 가중치 데이터(예를 들면, 커널)를 연산하도록 구성된 복수의 프로세싱 엘리먼트들(PE1…)(110)을 포함하도록 구성된다. 각각의 프로세싱 엘리먼트는 MAC (multiply and accumulate) 연산기 및/또는 ALU (Arithmetic Logic Unit) 연산기를 포함할 수 있다. 단, 본 개시에 따른 예시들은 이에 제한되지 않는다.

[0150] 도 2a에서는 예시적으로 복수의 프로세싱 엘리먼트들(PE1…)(110)이 도시되었지만, 하나의 프로세싱 엘리먼트 내부에 MAC을 대체하여, 복수의 곱셈기(multiplier) 및 가산기 트리(adder tree)로 구현된 연산기들이 병렬로 배치되어 구성되는 것도 가능하다. 이러한 경우, 프로세싱 엘리먼트 어레이(100)는 복수의 연산기를 포함하는 적어도 하나의 프로세싱 엘리먼트로 지칭되는 것도 가능하다.

[0151] 또한, 도 2a에 도시된 복수의 프로세싱 엘리먼트들(PE1…)(110)은 단지 설명의 편의를 위한 예시이며, 복수의

프로세싱 엘리먼트들(PE1…)(110)의 개수는 제한되지 않는다. 복수의 프로세싱 엘리먼트들(PE1…)(110)의 개수에 의해서 프로세싱 엘리먼트 어레이의 크기 또는 개수가 결정될 수 있다. 프로세싱 엘리먼트 어레이의 크기는 N x M 행렬 형태로 구현될 수 있다. 여기서 N과 M은 0보다 큰 정수이다. 이에, 프로세싱 엘리먼트 어레이(100)는 N x M 개의 프로세싱 엘리먼트를 포함할 수 있다. 즉, 프로세싱 엘리먼트는 1개 이상일 수 있다.

[0152] 또한, 프로세싱 엘리먼트 어레이(100)는 복수 서브 모듈로 구성되는 것도 가능하다. 이에, 프로세싱 엘리먼트 어레이(100)는 N x M x L 개의 서브 모듈로 구성된 프로세싱 엘리먼트를 포함할 수 있다. 부연 설명하면 L개는 프로세싱 엘리먼트 어레이의 서브 모듈의 개수로, 코어, 엔진 또는 쓰레드 등으로 지칭될 수 있다.

[0153] 프로세싱 엘리먼트 어레이(100)의 크기는 NPU(1000)가 작동하는 인공신경망모델의 특성을 고려하여 설계할 수 있다. 부연 설명하면, 프로세싱 엘리먼트의 개수는 작동할 인공신경망모델의 데이터 크기, 요구되는 동작 속도, 요구되는 소비 전력 등을 고려하여 결정될 수 있다. 인공신경망모델의 데이터 크기는 인공신경망모델의 레이어 수와 각각의 레이어의 가중치 데이터 크기에 대응되어 크기가 결정될 수 있다.

[0154] 따라서, 본 개시의 일 예시에 따른 프로세싱 엘리먼트 어레이(100)의 크기는 제한되지 않는다. 프로세싱 엘리먼트 어레이(100)의 프로세싱 엘리먼트들(PE1…)(110)의 개수가 증가할수록 작동하는 인공신경망모델의 병렬 연산 능력이 증가되나, NPU(1000)의 제조 비용 및 물리적인 크기가 증가될 수 있다.

[0155] 예를 들면, NPU(1000)에서 작동되는 인공신경망모델은 30개의 특정 키워드를 감지하도록 학습된 인공신경망, 즉 AI 키워드 인식모델일 수 있다. 이러한 경우, 엘리먼트 어레이(100)의 크기는 인공신경망모델의 연산량 특성을 고려하여 N x M로 설계될 수 있다. 다르게 설명하면, 엘리먼트 어레이(100)는 12개의 프로세싱 엘리먼트들을 포함할 수 있다. 단, 이에 제한되지 않으며, 복수의 프로세싱 엘리먼트들(PE1…)(110)의 개수는 예를 들면, 8개 내지 16,384 범위 내에서 선택되는 것도 가능하다. 즉, 본 개시의 예시들에서 프로세싱 엘리먼트의 개수는 제한되지 않는다.

[0156] 프로세싱 엘리먼트 어레이(100)는 인공신경망 연산에 필요한 덧셈, 곱셈, 누산 등의 기능을 수행하도록 구성된다. 다르게 설명하면, 프로세싱 엘리먼트 어레이(100)는 MAC(multiplication and accumulation) 연산을 수행하도록 구성될 수 있다.

[0157] 내부 메모리(200)는 휘발성 메모리일 수 있다. 휘발성 메모리는 전원이 공급된 경우에만 데이터를 저장하고, 전원 공급이 차단되면 저장된 데이터가 소멸되는 메모리일 수 있다. 휘발성 메모리는 정적 랜덤 액세스 메모리(Static Random Access Memory; SRAM), 동적 랜덤 액세스 메모리 (Dynamic Random Access Memory; DRAM) 등을 포함할 수 있다. 내부 메모리(200)는 바람직하게는 SRAM일 수 있으나, 이에 한정되지 않는다.

[0158] 이하에서는 인공신경망 중에서 심층 신경망(DNN, Deep Neural Network)의 한 종류인 컨볼루션 신경망(CNN, Convolutional Neural Network)에 대해서 집중적으로 설명하기로 한다.

[0159] 컨볼루션 신경망은 하나 또는 여러 개의 컨볼루션 레이어(convolutional layer)과 통합 레이어(pooling layer), 완전하게 연결된 레이어(fully connected layer)들의 조합일 수 있다. 컨볼루션 신경망은 2차원 데이터의 학습 및 추론에 적합한 구조를 가지고 있으며, 역전달(Backpropagation algorithm)을 통해 학습될 수 있다.

[0160] 본 개시의 예시에서, 컨볼루션 신경망에는 레이어마다 복수의 채널을 포함한다. 채널마다 채널의 입력 영상의 특징을 추출하는 커널이 존재한다. 커널은 2차원 행렬로 구성될 수 있으며, 입력 데이터를 순회하면서 컨볼루션 연산 수행한다. 커널의 크기(N x M)는 임의로 결정될 수 있으며, 커널이 입력 데이터를 순회하는 간격(stride) 또한 임의로 결정될 수 있다. 커널 하나당 입력 데이터 전체에 대한 커널의 일치 정도는 특징맵(feature map) 또는 활성화 맵 일 수 있다. 이하에서 커널은 일 세트의 가중치 값들 또는 복수의 세트의 가중치 값들을 포함할 수 있다.

[0161] 프로세싱 엘리먼트 어레이(100)는 인공신경망의 컨벌루션 연산을 처리하도록 구성되고, 활성화 함수 연산은 별도의 활성화 함수 처리 모듈에서 처리하도록 구성될 수 있다. 이러한 경우, 프로세싱 엘리먼트 어레이(100)는 컨벌루션 연산만을 위해서 동작될 수 있다. 특히 이러한 경우, 프로세싱 엘리먼트 어레이(100)는 정수 타입의 데이터만 처리하도록 구성되어, 방대한 합성곱 연산 시 연산 효율을 극대화하는 것도 가능하다.

[0162] 이처럼 컨볼루션 연산은 입력 데이터와 커널의 조합으로 이루어진 연산이므로, 이후 비선형성을 추가하기 위한 활성화 함수가 적용될 수 있다. 컨볼루션 연산의 결과인 특징맵에 활성화 함수가 적용되면 활성화 맵으로 지칭될 수 있다.

[0163] 일반적인 컨볼루션 신경망은 Alexnet, Squeezenet, VGG16, Resnet152, Moblienet 등이 있는데, 각 인공신경망

모델은 한번의 추론을 위해 각각 727 MFLOPs(Mega Floating-point Operations per Second), 837 MFLOPs, 16 MFLOPs, 11 MFLOPs, 11 MFLOPs, 579 MFLOPs의 곱셈 횟수가 필요하고, 커널을 포함한 모든 가중치가 각각 233 MB, 5 MB, 528 MB, 230 MB, 16 MB의 저장 크기를 갖는다. 따라서, 이러한 컨볼루션 신경망은 연산을 위해 상당히 많은 양의 하드웨어 리소스와 전력 소모량을 요구함을 알 수 있다.

[0164] 활성화 함수 적용을 위해서 프로세싱 엘리먼트 어레이(100)와 내부 메모리(200) 사이에 활성화 함수 처리 유닛이 더 배치될 수 있다. 활성화 함수 처리 유닛은 복수의 서브 모듈을 포함하도록 구성될 수 있다. 예를 들면 활성화 함수 처리 유닛은 ReLU 유닛, Leaky-ReLU 유닛, ReLU6 유닛, Swish 유닛, Sigmoid 유닛 Average Pooling 유닛, Skip connection 유닛, Squeeze and Excitation 유닛, Bias 유닛, Quantization 유닛, Dequantization 유닛, 하이퍼볼릭 탄젠트 유닛, Maxout 유닛, ELU 유닛, Batch-Normalization 유닛 중 복수개를 포함할 수 있다. 활성화 함수 처리 유닛은 각각의 서브 모듈을 파이프라인 구조로 배치하도록 구성될 수 있다.

[0165] 활성화 함수 처리 유닛은 각각의 서브 모듈을 선택적으로 활성화하거나 비활성화할 수 있다.

[0166] NPU 스케줄러(300)는 활성화 함수 처리 유닛을 제어하도록 구성될 수 있다.

[0167] NPU 스케줄러(300)는 인공신경망모델의 데이터 지역성에 기초하여 활성화 함수 처리 유닛의 각각의 서브 모듈을 선택적으로 활성화하거나 비활성화할 수 있다.

[0168] 활성화 함수 처리 유닛은 프로세싱 엘리먼트 어레이(100)에서 출력하는 각각의 배치채널의 특징맵을 순차적으로 처리하여 각각의 배치채널의 활성화맵을 출력하도록 구성될 수 있다.

[0169] 이에 대해서 구체적으로 도 2b를 참조하여 설명하도록 한다.

[0170] 도 2b는 NPU(1000)의 동작시 소모되는 에너지를 나타낸 예시도이다. 제시된 예시에서는 후술할 도 2c의 제 1 프로세싱 엘리먼트(PE1)(110)의 구성들(예: 곱셈기(641), 및 가산기(642))을 참조하여 설명하도록 한다.

[0171] 도 2b를 참조하면, 에너지 소모는 메모리 액세스, 덧셈 연산 및 곱셈 연산으로 구분될 수 있다.

[0172] 도 2b는 NPU(1000)의 동작시 소모되는 에너지를 나타낸 예시도이다.

[0173] "8b Add"는 가산기(642)의 8비트 정수 덧셈 연산을 의미한다. 8비트 정수 덧셈 연산은 0.03pj의 에너지를 소비할 수 있다.

[0174] "16b Add"는 가산기(642)의 16비트 정수 덧셈 연산을 의미한다. 16비트 정수 덧셈 연산은 0.05pj의 에너지를 소비할 수 있다.

[0175] "32b Add"는 가산기(642)의 32비트 정수 덧셈 연산을 의미한다. 32비트 정수 덧셈 연산은 0.1pj의 에너지를 소비할 수 있다.

[0176] "16b FP Add"는 가산기(642)의 16비트 부동소수점 덧셈 연산을 의미한다. 16비트 부동소수점 덧셈 연산은 0.4pj의 에너지를 소비할 수 있다.

[0177] "32b FP Add"는 가산기(642)의 32비트 부동소수점 덧셈 연산을 의미한다. 32비트 부동소수점 덧셈 연산은 0.9pj의 에너지를 소비할 수 있다.

[0178] "8b Mult"는 곱셈기(641)의 8비트 정수 곱셈 연산을 의미한다. 8비트 정수 곱셈 연산은 0.2pj의 에너지를 소비할 수 있다.

[0179] "32b Mult"는 곱셈기(641)의 32비트 정수 곱셈 연산을 의미한다. 32비트 정수 곱셈 연산은 3.1pj의 에너지를 소비할 수 있다.

[0180] "16b FP Mult"는 곱셈기(641)의 16비트 부동소수점 곱셈 연산을 의미한다. 16비트 부동소수점 곱셈 연산은 1.1pj의 에너지를 소비할 수 있다.

[0181] "32b FP Mult"는 곱셈기(641)의 32비트 부동소수점 곱셈 연산을 의미한다. 32비트 부동소수점 곱셈 연산은 3.7pj의 에너지를 소비할 수 있다.

[0182] "32b SRAM Read"는 NPU 메모리 시스템의 내부 메모리가 SRAM(static random access memory)일 경우, 32비트의 데이터 읽기 액세스를 의미한다. 32비트의 데이터를 NPU 메모리 시스템에서 읽어오는데 5pj의 에너지를 소비할 수 있다.

[0183] "32b DRAM Read"는 차량 제어 장치의 저장부가 DRAM일 경우, 32비트의 데이터 읽기 액세스를 의미한다. 32비트 데이터를 저장부에서 NPU 메모리 시스템으로 읽어오는데 640pj의 에너지를 소비할 수 있다. 에너지 단위는 피코-줄(pj)을 의미한다.

[0184] 종래의 신경 프로세싱 유닛은 이러한 커널들을 대응되는 채널마다 메모리에 저장하고, 컨볼루션 과정마다 메모리에서 불러와 입력 데이터를 처리하였다. 예를 들면, 컨볼루션 과정의 32비트 읽기 동작에서, NPU(1000)의 내부 메모리인 SRAM은 도 2b에 도시된 바와 같이 5pj의 전력을 소모하고, 메인 메모리인 DRAM은 640pj의 전력을 소모하였다. 이러한 메모리들은 8비트 덧셈 연산에서 0.03pj의 전력을, 16비트 덧셈에서 0.05pj의 전력을, 32비트 덧셈에서 0.1pj의 전력을, 8비트 곱셈에서 0.2pj의 전력을 소비하였다. 이처럼 종래의 신경 프로세싱 유닛은 소모되는 다른 연산에 비해 상당히 많은 전력을 소모하여 전체적인 성능 저하를 야기하는 문제가 있었다. 즉, NPU(1000)의 메인 메모리에서 커널을 읽을 때 소비되는 전력은 내부 메모리에서 커널을 읽을 때 소비되는 전력에 비해 128배 더 많이 소비되었다.

[0185] 즉, 메인 메모리(4000)의 동작 속도는 내부 메모리(200) 대비 느린데 비해, 단위 연산 당 전력 소모량은 상대적으로 훨씬 더 크기 때문에, 메인 메모리(4000)의 읽기 동작을 최소화하는 것이 NPU(1000)의 소비 전력 저감에 영향을 줄 수 있다. 특히 복수의 채널을 개별적으로 처리하면 소비 전력 효율이 특히 더 저하될 수 있다.

[0186] 이러한 비효율성을 극복하기 위해, 본 개시는 일 세트의 가중치 값들 또는 복수의 배치채널들을 불러오는 메인 메모리(4000)와 온칩 영역(A) 사이의 데이터 이동을 최소화하여 전체적인 하드웨어 리소스와 데이터 이동에 따른 전력 소모를 줄이고, 개선된 연산 성능을 갖는 신경 프로세싱 유닛을 제안한다.

[0187] 신경 프로세싱 유닛에서 객체 인식 모델을 이용하여 객체 인식을 수행하기 위해 배치채널들을 입력으로 이용하는 것은 객체 인식 모델의 가중치 값을 DRAM에서 액세스하는 횟수를 최소화하기 위한 것이다. 배치 데이터의 개수가 증가할수록 DRAM에 저장된 가중치 값에 액세스하는 횟수가 증가한다. 즉, 배치채널들의 개수에 비례하여 DRAM에 저장된 가중치 값에 액세스하는 횟수가 증가할 수 있다.

[0188] 따라서, 본 발명은 객체 인식을 위해 이용되는 객체 인식 모델에 관한 데이터를 SRAM으로 구성된 NPU 내부 메모리에 저장함으로써, NPU의 단위 동작 당 에너지 소모를 줄여 NPU의 성능이 보다 향상시킬 수 있다.

[0189] 이를 통해서 본 개시의 NPU가 장착된 자율 주행 차량은 차량의 안전한 자율 주행을 위해 지속적으로 인식해야 하는 전후좌우 접근 차량, 장애물, 신호등의 신호 정보 및 보행자 등과 같은 목표 객체를 인식하는데 소요되는 시간 및 객체 인식을 위해 소비되는 자원량을 최소화할 수 있다.

[0190] 이하에서는 신경 프로세싱 유닛이 전체적인 하드웨어 리소스와 데이터 이동에 따른 전력 소모를 줄이고, 개선된 연산 성능을 갖도록 하기 위해 프로세싱 엘리먼트 어레이 중 도 2a의 제 1 프로세싱 엘리먼트(PE1)를 예를 들어 설명한다.

[0191] 도 2c는 본 개시에 적용될 수 있는 프로세싱 엘리먼트 어레이 중 하나의 프로세싱 엘리먼트를 설명하는 개략적인 개념도이다.

[0192] 본 개시의 일 예시에 따른 NPU(1000)는 프로세싱 엘리먼트 어레이(100), 프로세싱 엘리먼트 어레이(100)에서 추론될 수 있는 인공신경망모델을 저장하도록 구성된 NPU 내부 메모리(200) 및 프로세싱 엘리먼트 어레이(100) 및 NPU 내부 메모리(200)를 제어하도록 구성된 NPU 스케줄러(300)를 포함한다. 프로세싱 엘리먼트 어레이(100)는 MAC 연산을 수행하도록 구성되고, MAC 연산 결과를 양자화해서 출력하도록 구성될 수 있다. 단, 본 개시의 예시들은 이에 제한되지 않는다.

[0193] NPU 내부 메모리(200)는 메모리 크기와 인공신경망모델의 데이터 크기에 따라 인공신경망모델의 전부 또는 일부를 저장할 수 있다.

[0194] 도 2c를 참조하면, 제 1 프로세싱 엘리먼트(PE1)(110)는 곱셈기(Multiplier)(641), 가산기(Adder)(642), 및 누산기(Accumulator)(643)를 포함할 수 있다. 단, 본 개시에 따른 예시들은 이에 제한되지 않으며, 프로세싱 엘리먼트 어레이(100)는 인공신경망의 연산 특성을 고려하여 변형 실시될 수도 있다.

[0195] 곱셈기(641)는 입력받은 (N)bit 데이터와 (M)bit 데이터를 곱한다. 곱셈기(641)의 연산 값은 (N+M)bit 데이터로 출력된다. 여기서 N과 M은 0보다 큰 정수이다. (N)bit 데이터를 입력받는 제 1 입력부는 변수 같은 특성을 가지는 값을 입력 받도록 구성될 수 있고, (M)bit 데이터를 입력 받는 제 2 입력부는 상수 같은 특성을 가지는 값을 입력 받도록 구성될 수 있다.

[0196] 여기서 변수 같은 특성을 가지는 값 또는 변수는, 해당 값이 저장된 메모리 어드레스의 값인 경우, 들어오는 입력 데이터가 갱신될 때마다 갱신되는 값을 의미한다. 예를 들면, 각 레이어의 노드 데이터는 인공신경망모델의 가중치 데이터가 반영된 MAC 연산 값일 수 있으며, 해당 인공신경망모델로 동영상 데이터의 객체 인식 등을 추론할 경우, 매 프레임마다 입력 영상이 바뀌기 때문에, 각 레이어의 노드 데이터는 변하게 된다.

[0197] 여기서 상수 같은 특성을 가지는 값 또는 상수는, 해당 값이 저장된 메모리 어드레스의 값인 경우, 들어오는 입력 데이터의 갱신과 상관없이 보존되는 값을 의미한다. 예를 들면, 연결망의 가중치 데이터는 인공신경망모델의 고유한 추론 판단 기준이며, 해당 인공신경망모델로 동영상 데이터의 객체 인식 등을 추론하더라도, 연결망의 가중치 데이터는 변하지 않을 수 있다.

[0198] 즉, 곱셈기(641)는 하나의 변수와 하나의 상수를 입력 받도록 구성될 수 있다. 부연 설명하면, 제 1 입력부에 입력되는 변수 값은 인공신경망의 레이어의 노드 데이터일 수 있으며, 노드 데이터는 인공신경망의 입력 레이어의 입력 데이터, 은닉 레이어의 누산 값, 및 출력 레이어의 누산 값일 수 있다. 제 2 입력부에 입력되는 상수 값은 인공신경망의 연결망의 가중치 데이터일 수 있다.

[0199] 이처럼 NPU 스케줄러(300)가 변수 값과 상수 값의 특성을 구분할 경우, NPU 스케줄러(300)는 NPU 내부 메모리(200)의 메모리 재사용율을 증가시킬 수 있다. 단, 곱셈기(641)의 입력 데이터는 상수 값과 변수 값에 제한되지 않는다. 즉, 본 개시의 예시들에 따르면, 프로세싱 엘리먼트의 입력 데이터는 상수 값과 변수 값의 특성을 이해하여 동작할 수 있기 때문에, NPU(1000)의 연산 효율을 향상시킬 수 있다. 하지만 NPU(1000)의 동작은 입력 데이터의 상수 값 및 변수 값의 특징에 제한되지 않는다.

[0200] 이를 바탕으로, NPU 스케줄러(300)는 상수 값의 특성을 고려하여 메모리 재사용율을 향상시키도록 구성될 수 있다.

[0201] 변수 값은 각 레이어의 연산 값이며, NPU 스케줄러(300)는 인공신경망모델의 데이터 지역성 정보 또는 구조에 대한 정보에 기초하여 재사용 가능한 변수 값을 인식하고, 메모리를 재사용 하도록 NPU 내부 메모리(200)를 제어할 수 있다.

[0202] 상수 값은 각 연결망의 가중치 데이터이므로, NPU 스케줄러(300)는 인공신경망모델의 데이터 지역성 정보 또는 구조에 대한 정보에 기초하여 반복 사용되는 연결망의 상수 값을 인식하고, 메모리를 재사용 하도록 NPU 내부 메모리(200)를 제어할 수 있다.

[0203] 즉, NPU 스케줄러(300)는 인공신경망모델의 데이터 지역성 정보 또는 구조에 대한 정보에 기초하여 재사용 가능한 변수 값 및 재사용 가능한 상수 값을 인식하고, NPU 스케줄러(300)는 메모리를 재사용 하도록 NPU 내부 메모리(200)를 제어하도록 구성될 수 있다.

[0204] 한편, 제 1 프로세싱 엘리먼트(PE1)(110)는 곱셈기(641)의 제 1 입력부 및 제 2 입력부 중 하나의 입력부에 0이 입력될 때, 연산을 하지 않더라도 연산 결과가 0인 것을 인지하고 있기 때문에, 곱셈기(641)가 연산을 하지 않도록 동작을 제한할 수 있다.

[0205] 예를 들면, 곱셈기(641)의 제 1 입력부 및 제 2 입력부 중 하나의 입력부에 0이 입력될 때, 곱셈기(641)는 제로 스키핑(zero skipping) 방식으로 동작하도록 구성될 수 있다.

[0206] 곱셈기(641)의 제 1 입력부 및 제 2 입력부에 입력되는 데이터는 인공신경망모델의 각각의 레이어의 노드 데이터 및 가중치 데이터의 양자화에 따라서 비트 폭(bit width)이 결정될 수 있다. 예를 들면, 제 1 레이어의 노드 데이터가 5bit로 양자화 되고 제 1 레이어의 가중치 데이터가 7bit로 양자화되는 경우 제 1 입력부는 5bit의 데이터를 입력 받도록 구성되고, 제 2 입력부는 7bit의 데이터를 입력 받도록 구성될 수 있다.

[0207] NPU(1000)는 NPU 내부 메모리(200)에 저장된 양자화된 데이터가 제 1 프로세싱 엘리먼트(PE1)(110)의 입력부들에 입력될 때 양자화된 비트 폭이 실시간으로 변환되도록 제 1 프로세싱 엘리먼트(PE1)(110)를 제어할 수 있다. 즉, 레이어 마다 양자화 된 비트 폭이 다를 수 있으므로, 제 1 프로세싱 엘리먼트(PE1)(110)는 입력되는 데이터의 비트 폭이 변환될 때 실시간으로 비트 폭 정보를 NPU(1000)에서 제공받고, 제공된 비트 폭 정보에 기반하여 실시간으로 비트 폭을 변환시켜서 입력 데이터를 생성하도록 구성될 수 있다.

[0208] 가산기(642)는 곱셈기(641)의 연산 값과 누산기(643)의 연산 값을 가산한다. (L)loops가 0일 경우, 누산된 데이터가 없으므로, 가산기(642)의 연산 값은 곱셈기(111)의 연산 값과 동일할 수 있다. (L)loops가 1일 경우, 곱셈기(641)의 연산 값과 누산기(643)의 연산 값이 가산된 값이 가산기의 연산 값일 수 있다.

[0209] 누산기(643)는 가산기(642)의 연산 값과 곱셈기(641)의 연산 값이 (L)loops 횟수만큼 누산되도록 가산기(642)의 출력부에서 출력된 데이터를 임시 저장한다. 구체적으로, 가산기(642)의 출력부에서 출력된 가산기(642)의 연산 값은 누산기(643)의 입력부에 입력되고, 입력된 연산 값은 누산기(643)에 임시 저장되었다가 누산기(643)의 출력부에서 출력된다. 출력된 연산 값은 루프에 의해 가산기(642)의 입력부에 입력된다. 이때, 가산기의 입력부(642)에는 곱셈기(641)의 출력부에서 새롭게 출력된 연산 값이 함께 입력된다. 즉, 누산기(643)의 연산 값과 곱셈기(641)의 새로운 연산 값이 가산기(642)의 입력부에 입력되고, 이 값들이 가산기(642)에서 가산되어 가산기(642)의 출력부를 통해 출력된다. 가산기(642)의 출력부에서 출력된 데이터, 즉 가산기(642)의 새로운 연산 값은 누산기(643)의 입력부에 입력되며, 이후 동작들은 상술한 동작들과 실질적으로 동일하게 루프 횟수만큼 수행된다.

[0210] 이처럼, 누산기(643)는 곱셈기(641)의 연산 값과 가산기(642)의 연산 값을 루프 횟수만큼 누산하기 위해 가산기(642)의 출력부에서 출력된 데이터를 임시 저장하므로, 누산기(643)의 입력부에 입력되는 데이터 및 출력부에서 출력되는 데이터는 가산기(642)의 출력부에서 출력된 데이터와 같은 (N+M+log2(L))bit의 비트 폭을 가질 수 있다. 여기서 L은 0보다 큰 정수이다.

[0211] 누산기(643)는 누산이 종료되면, 초기화 신호(initialization reset)를 인가받아서 누산기(643) 내부에 저장된 데이터를 0으로 초기화 할 수 있다. 단, 본 개시에 따른 예시들은 이에 제한되지 않는다.

[0212] 누산기(643)의 출력 데이터(N+M+log2(L))bit는 다음 레이어의 노드 데이터 또는 컨볼루션의 입력 데이터가 될 수 있다.

[0213] 다양한 실시예에서 제 1 프로세싱 엘리먼트(PE1)(110)는 비트 양자화 유닛을 더 포함할 수 있다. 예를 들어, 비트 양자화 유닛은 누산기(643)에서 출력되는 데이터의 비트 폭을 저감할 수 있다. 비트 양자화 유닛은 NPU 스케줄러(300)에 의해서 제어될 수 있다. 양자화된 데이터의 비트 폭은 (X)bit로 출력될 수 있다. 여기서 X는 0보다 큰 정수이다. 상술한 구성에 따르면, 프로세싱 엘리먼트 어레이(110)는 MAC 연산을 수행하도록 구성되고 MAC 연산 결과를 양자화해서 출력할 수 있다. 특히 이러한 양자화는 (L)loops가 증가할수록 소비 전력을 더 절감할 수 있는 효과가 있다. 또한 소비 전력이 저감되면 발열도 저감할 수 있다. 특히 발열을 저감하면 NPU(1000)의 고온에 의한 오동작 발생 가능성을 저감할 수 있다.

[0214] 비트 양자화 유닛의 출력 데이터(X)bit는 다음 레이어의 노드 데이터 또는 컨볼루션의 입력 데이터가 될 수 있다. 만약 인공신경망모델이 양자화되었다면, 비트 양자화 유닛은 양자화된 정보를 인공신경망모델에서 제공받도록 구성될 수 있다. 단, 이에 제한되지 않으며, NPU 스케줄러(300)는 인공신경망모델을 분석하여 양자화된 정보를 추출하도록 구성될 수 있다. 따라서 비트 양자화 유닛은 양자화된 데이터 크기에 대응되도록, 출력 데이터(X)bit를 양자화 된 비트 폭으로 변환하여 출력할 수 있다. 비트 양자화 유닛의 출력 데이터(X)bit는 양자화된 비트 폭으로 NPU 내부 메모리(200)에 저장될 수 있다. 비트 양자화 유닛은 프로세싱 엘리먼트 또는 활성화 함수 처리 유닛에 포함될 수 있다.

[0215] 본 개시의 일 예시에 따른 NPU(1000)의 프로세싱 엘리먼트 어레이(110)는 비트 양자화 유닛에 의해서 누산기(643)에서 출력되는 (N+M+log2(L))bit의 비트 폭의 데이터를 (X)bit의 비트 폭으로 저감할 수 있다. 이를 위해 NPU 스케줄러(300)는 비트 양자화 유닛을 제어하여 출력 데이터의 비트 폭을 LSB(least significant bit)에서 MSB(most significant bit)까지 소정 비트만큼 저감할 수 있다. 출력 데이터의 비트 폭이 저감되면 NPU(1000)의 소비 전력, 연산량, 메모리 사용량이 저감될 수 있다. 하지만 비트 폭이 특정 길이 이하로 저감될 경우, 인공신경망모델의 추론 정확도가 급격히 저하될 수 있는 문제가 발생될 수 있다. 따라서, 출력 데이터의 비트 폭 저감, 즉, 양자화 수준은 인공신경망모델의 추론 정확도 저감 수준 대비 소비 전력, 연산량, 메모리 사용량 저감 정도를 비교하여 결정될 수 있다. 양자화 수준은 인공신경망모델의 목표 추론 정확도를 결정하고, 비트 폭을 점진적으로 저감하면서 테스트하는 방법으로 결정될 수 있다. 양자화 수준은 각각의 레이어의 연산 값마다 각각 결정될 수 있다.

[0216] 상술한 제 1 프로세싱 엘리먼트(PE1)에 따라 곱셈기(641)의 (N)bit 데이터와 (M)bit 데이터의 비트 폭을 조절하고, 비트 양자화 유닛에 의해서 연산 값(X)bit의 비트 폭을 저감함으로써, 프로세싱 엘리먼트 어레이의 MAC 연산 속도를 향상시키면서 소비 전력을 저감할 수 있고, 인공신경망의 컨볼루션(convolution) 연산을 보다 더 효율적으로 할 수 있다.

[0217] 단, 본 개시의 비트 양자화 유닛은 프로세싱 엘리먼트가 아닌 활성화 함수 처리 유닛에 포함되도록 구성되는 것도 가능하다.

[0218] 이를 바탕으로 NPU(1000)의 NPU 내부 메모리(200)는 프로세싱 엘리먼트 어레이(100)의 MAC 연산 특성 및 소비 전력 특성을 고려하여 구성된 메모리 시스템일 수 있다.

[0219] 예를 들면, NPU(1000)는, 프로세싱 엘리먼트 어레이(100)의 MAC 연산 특성 및 소비 전력 특성을 고려하여 프로세싱 엘리먼트 어레이(100)의 연산 값의 비트 폭을 저감하도록 구성될 수 있다.

[0220] NPU(1000)의 NPU 내부 메모리(200)는 NPU(1000)의 소비 전력을 최소화하도록 구성될 수 있다.

[0221] NPU(1000)의 NPU 내부 메모리(200)는 작동되는 인공신경망모델의 파라미터들의 크기 및 연산 단계를 고려하여 저전력으로 메모리를 제어하도록 구성된 메모리 시스템일 수 있다.

[0222] NPU(1000)의 NPU 내부 메모리(200)는 인공신경망모델의 데이터 크기 및 연산 단계를 고려하여 가중치 데이터가 저장된 특정 메모리 어드레스를 재사용하도록 구성된 저전력 메모리 시스템일 수 있다.

[0223] NPU(1000)는 비선형성을 부여하기 위한 여러 가지 활성화 함수를 제공할 수 있다. 예를 들면, 활성화 함수는 입력값에 대한 비선형의 출력값을 도출하는 시그모이드 함수, 하이퍼볼릭 탄젠트(tanh) 함수, ReLU함수, Leaky ReLU 함수, Maxout 함수 또는 ELU 함수 등을 포함할 수 있으나, 이에 한정되지 않는다. 이러한 활성화 함수는 MAC 연산 이후에 선택적으로 적용될 수 있다. 활성화 함수가 적용된 연산 값은, 활성화 맵으로 지칭될 수 있다. 활성화 함수 적용 전의 연산 값은 특징맵으로 지칭될 수 있다.

[0224] 도 3은 도 2a에 도시된 NPU(1000)의 변형예를 나타낸 예시도이다.

[0225] 도 3에 도시된 NPU(1000)는 도 2a에 예시적으로 도시된 프로세싱 유닛(1000)과 비교하면, 프로세싱 엘리먼트(110')를 제외하곤 실질적으로 동일하기 때문에, 이하 단지 설명의 편의를 위해서 중복 설명은 생략할 수 있다.

[0226] 도 3에 예시적으로 도시된 프로세싱 엘리먼트 어레이(100)는 복수의 프로세싱 엘리먼트들(PE1….) 외에, 각각의 프로세싱 엘리먼트들(PE1…)에 대응되는 각각의 레지스터 파일들(RF1…)을 더 포함할 수 있다.

[0227] 도 3에 도시된 복수의 프로세싱 엘리먼트들(PE1…) 및 복수의 레지스터 파일들(RF1…)은 단지 설명의 편의를 위한 예시이며, 복수의 프로세싱 엘리먼트들(PE1…) 및 복수의 레지스터 파일들(RF1…)의 개수는 제한되지 않는다.

[0228] 프로세싱 엘리먼트 어레이(100)의 크기 또는 개수는 복수의 프로세싱 엘리먼트들(PE1…) 및 복수의 레지스터 파일들(RF1...)의 개수에 의해서 결정될 수 있다. 프로세싱 엘리먼트 어레이(100) 및 복수의 레지스터 파일들(RF1...)의 크기는 N x M 행렬 형태로 구현될 수 있다. 여기서 N 과 M은 0보다 큰 정수이다.

[0229] 프로세싱 엘리먼트 어레이(100)의 어레이 크기는 NPU(1000)가 작동하는 인공신경망모델의 특성을 고려하여 설계될 수 있다. 부연 설명하면, 레지스터 파일의 메모리 크기는 작동할 인공신경망모델의 데이터 크기, 요구되는 동작 속도, 요구되는 소비 전력 등을 고려하여 결정될 수 있다.

[0230] 프로세싱 엘리먼트 어레이(100)의 레지스터 파일들(RF1...)은 프로세싱 엘리먼트들(PE1 to PE12)과 직접 연결된 정적 메모리 유닛이다. 레지스터 파일들(RF1...)은 예를 들면, 플립플롭, 및/또는 래치 등으로 구성될 수 있다. 레지스터 파일들(RF1...)은 대응되는 프로세싱 엘리먼트들(RF1...)의 MAC 연산 값을 저장하도록 구성될 수 있다. 레지스터 파일들(RF1...)은 NPU 시스템 메모리(200)와 가중치 데이터 및/또는 노드 데이터를 제공하거나 제공받도록 구성될 수 있다. 레지스터 파일들(RF1…)은 누산기의 기능을 수행하도록 구성되는 것도 가능하다.

[0231] 활성화 함수 적용을 위해서 프로세싱 엘리먼트 어레이(100)와 내부 메모리(200) 사이에 활성화 함수 처리 유닛이 더 배치될 수 있다.

[0232] 도 4는 예시적인 인공신경망모델을 설명하는 개략적인 개념도이다.

[0233] 도 4를 참조하면, 컨볼루션 신경망은 적어도 하나의 컨볼루션 레이어, 적어도 하나의 풀링 레이어, 및 적어도 하나의 완전 연결 레이어를 포함한다.

[0234] 예를 들면, 컨볼루션은, 입력 데이터의 크기(통상적으로 1×1, 3×3 또는 5×5 행렬)와 출력 특징맵(Feature Map)의 깊이(커널의 수)와 같은 두 개의 주요 파라미터에 의해 정의될 수 있다. 이러한 주요 파라미터는 컨볼루션에 의해 연산될 수 있다. 이들 컨볼루션은, 깊이 32에서 시작하여, 깊이 64로 계속되며, 깊이 128 또는 256에서 종료될 수 있다. 컨볼루션 연산은, 입력 데이터인 입력 이미지 행렬 위로 3×3 또는 5×5 크기의 커널(kernel)을 슬라이딩하여 커널의 각 원소와 겹쳐지는 입력 이미지 행렬의 각 원소를 곱한 후 이들을 모두 더하는 연산을 의미한다. 여기서, 입력 이미지 행렬은 3차원 패치(3D patch)이며, 커널은 가중치라고 하는 동일한

학습 가중치 행렬을 의미한다.

[0235] 다시 말해서, 컨볼루션은 3차원 패치가 학습 가중치 행렬과의 텐서 곱에 의해 1차원 벡터로 변환되고, 이러한 벡터가 3차원 출력 특징맵(feature map)으로 공간적으로 재조립되는 동작을 의미한다. 출력 특징맵의 모든 공간 위치는 입력 특징맵의 동일한 위치에 대응될 수 있다.

[0236] 컨볼루션 레이어는, 학습 과정동안 많은 그라디언트 업데이트 반복에 걸쳐 학습되는 커널(즉, 가중치 행렬)과 입력 데이터 간의 컨볼루션을 수행할 수 있다. (m, n)을 커널 크기라고 하고 W를 가중치 값이라고 설정하면, 컨볼루션 레이어는 내적을 계산함으로써 입력 데이터와 가중치 행렬의 컨볼루션을 수행할 수 있다.

[0237] 커널이 입력 데이터를 가로질러 슬라이딩하는 단차 크기를 간격이라고 하며, 커널 면적(m×n)을 수용장(receptive field)이라고 할 수 있다. 동일한 컨볼루션 커널이 입력의 상이한 위치에 걸쳐 적용되며, 이는 학습되는 커널의 수를 감소시킨다. 이것은, 또한, 위치 불변 학습을 가능하게 하며, 중요한 패턴이 입력에 존재하는 경우, 컨볼루션 필터(즉, 커널)는 시퀀스의 위치에 관계없이 그 패턴을 학습할 수 있다.

[0238] 이와 같이 생성된 출력 특징맵에 활성화 함수가 적용되어 활성화 맵이 최종적으로 출력될 수 있다. 또한, 현재 레이어에서의 사용된 가중치는 컨볼루션을 통해 다음 레이어에 전달될 수 있다. 풀링 레이어는 출력 데이터(즉, 활성화 맵)을 다운샘플링하여 특징맵의 크기를 줄이는 풀링 연산을 수행할 수 있다. 예를 들어, 풀링 연산은 최대 풀링(max pooling) 및/또는 평균 풀링(average pooling)을 포함할 수 있으나, 이에 한정되지 않는다. 최대 풀링 연산은 커널을 이용하며, 특징맵과 커널이 슬라이딩되어 커널과 겹쳐지는 특징맵의 영역에서 최대 값을 출력한다. 평균 풀링 연산은 특징맵과 커널이 슬라이딩되어 커널과 겹쳐지는 특징맵의 영역 내에서 평균값을 출력한다. 이처럼 풀링 연산에 의해 특징맵의 크기가 줄어들기 때문에 특징맵의 가중치 개수 또한 줄어든다.

[0239] 완전 연결 레이어는 풀링 레이어를 통해서 출력된 데이터를 복수의 클래스(즉, 추정값)로 분류하고, 분류된 클래스 및 이에 대한 점수(score)를 출력할 수 있다. 풀링 레이어를 통해서 출력된 데이터는 3차원 특징맵 형태를 이루며, 이러한 3차원 특징맵이 1차원 벡터로 변환되어 완전 연결 레이어로 입력될 수 있다.

[0240] 컨볼루션 신경망은, 입력 데이터가 특정 출력 추정값으로 이어지도록 조정되거나 학습될 수 있다. 다시 말해서, 컨볼루션 신경망은 출력 추정값이 실측 자료(ground truth)에 점진적으로 일치하거나 근접할 때까지 출력 추정값과 실측 자료 간의 비교에 기초하여 역전파(backpropagation)를 이용하여 조정될 수 있다.

[0241] 컨볼루션 신경망은, 실측 자료와 실제 출력 간의 차이에 기초하는 뉴런들 간의 가중치를 조정함으로써 학습될 수 있다.

[0242] 이하에서는 도 5 내지 도 12를 참조하여, 본 개시의 다양한 예시에 따른 신경 프로세싱 유닛이 ANN에 대한 동작들을 수행하는 방법 및 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 개략적으로 설명한다.

[0243] 이하에서 가중치 및 배치채널 A 내지 D가 참조되는데, 이들의 크기 및 분할은 예시적이며, 설명의 편의를 위해 서로 동일한 크기를 가지거나 상대적인 크기를 가지도록 도시된다. 또한, 이들은 메모리 공간에 할당되는 주소를 가지며, 도면에서 데이터의 이동은 메모리 주소에 다른 데이터가 쓰여지는 것을 의미할 수 있다. 같은 범주에서 동일한 위치의 데이터는 다른 데이터가 쓰이거나 하지 않고, 유지되도록 의도된다. 또한, 각 단계들은 한 클럭 또는 복수의 클럭 동안의 계산 시간을 의미할 수도 있으나 이에 제한되지 않고, 가변적 클럭 동안 수행될 수도 있으며, 각 단계가 동일한 클럭 동안 수행되는 것이 의도되지 않는다. 또한, 각 단계들은 메모리의 매우 짧은 시간 동안의 상태이며 정적으로 고정되는 상태가 아니라는 점을 유념해야 한다.

[0244] 게다가, 메모리의 공간들은 예시적으로 동일한 구획 또는 크기를 가지는 것으로 표시되지만, 제한되지 않고, 메모리의 공간들은 다양한 구획 (예를 들어, 조각난 구획) 을 가질 수도 있고 다양한 크기를 가질 수도 있다. 또한 본 예시에서 S는 step을 의미할 수 있다.

[0245] 본 예시에서 ANN은 복수의 배치채널들로부터의 객체 검출, 분류 또는 세그먼트화를 포함하는 적어도 하나의 동작을 수행하도록 구성된다. ANN에 대한 동작들 전에 복수의 배치채널들이 전처리될 수 있으며, 복수의 배치채널들 각각은 복수의 이미지들 각각에 대응한다.

[0246] 도 5는 본 개시의 일 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다. 도 6은 본 개시의 일 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

[0247] 제시된 예시에서 복수의 배치채널은 제 1 배치채널(Batch A), 제 2 배치채널(Batch B), 제 3 배치채널(Batch

C), 및 제 4 배치채널(Batch D)을 포함한다. 각각의 배치채널은 예를 들어 4 개의 부분(또는 4 개의 일부)으로 나뉘어 질 수 있다. 이러한 배치채널들 각각은 완전한 데이터 세트를 포함할 수 있다.

[0248] 먼저, 도 5를 참조하면 일 세트의 가중치, 제 1 배치채널의 적어도 일부, 및 제 2 배치채널의 적어도 일부가 적어도 하나의 메모리에 저장된다(S2001). 여기서, 일 세트의 가중치는 적어도 하나의 가중치 값을 포함하는 가중치 행렬을 의미하고, 메모리는 내부 메모리(200), 온칩 메모리 또는 메인 메모리일 수 있다.

[0249] 본 예시에서 적어도 하나의 메모리에 저장되기 전에, 일 세트의 가중치의 크기, 제 1 배치채널의 적어도 일부의 크기, 및 제 2 배치채널의 적어도 일부의 크기는 적어도 하나의 메모리에 피팅되도록 조정될 수 있다.

[0250] 본 예시에서 제 1 배치채널의 적어도 일부의 크기는 적어도 하나의 메모리의 크기를 복수의 배치채널들의 수로 나눈 것과 같거나 작을 수 있다. 또한 적어도 하나의 내부 메모리의 크기는 ANN의 가장 큰 특징맵 크기 및 배치채널의 수에 대응할 수 있다. 본 예시에서 적어도 하나의 내부 메모리는 ANN의 압축된 파라미터들을 저장할 수 있다.

[0251] 부연 설명하면, 적어도 하나의 메모리의 크기는 신경 프로세싱 유닛(1000)이 처리하고자 하는 특정 ANN의 파라미터의 데이터 크기와 배치채널의 개수를 고려해서 결정될 수 있다.

[0252] 다음으로, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부 각각과 일 세트의 가중치가 계산된다(S2003). 해당 계산은 예를 들어 간격에 따른 컨볼루션 연산에 대응할 수 있다.

[0253] 다음으로, 일 세트의 가중치를 유지하면서 제 1 배치채널의 다음 일부 및 제 2 배치채널의 다음 일부가 적어도 하나의 메모리에 저장되고(S2005), 제 1 배치채널의 다음 일부 및 제 2 배치채널의 다음 일부와 일 세트의 가중치가 계산된다(S2007).

[0254] 이후 S2005 및 S2007 단계가 반복되면서 인공신경망 연산이 수행된다(S2009).

[0255] 이에 대해서 구체적으로 도 6을 참조하여 설명하면, 본 예시에서 메모리는 5개의 메모리 공간을 가진다고 가정한다. 또한 본 예시에서 일 세트의 가중치는 가중치(W)를 포함하고, 제 1 배치채널(Batch A)은 A1, A2, A3, A4를 포함하고, 제 2 배치채널(Batch B)은 B1, B2, B3, B4를 포함하고, 제 3 배치채널(Batch C)은 C1, C2, C3, C4를 포함하며, 제 4 배치채널(Batch D)은 D1, D2, D3, D4를 포함한다고 가정한다.

[0256] 도 6을 참조하면, S1에서 5개의 메모리 공간에는 가중치(W), 제 1 배치채널의 1부분(A1), 제 2 배치채널의 1부분(B1), 제 3 배치채널의 1부분(C1), 및 제 4 배치채널의 1부분(D1)이 채워진다. 프로세싱 엘리먼트(PE)는 가중치(W)와 제 1, 제 2, 제 3 및 제 4 배치채널들의 1부분들 각각에 대한 계산을 수행한다. 여기서, 가중치(W)는 적어도 하나의 가중치 값을 포함하는 가중치 행렬일 수 있다.

[0257] S2에서 가중치(W)는 S1에서와 같이 유지된 상태에서 제 1 배치채널의 2부분(A2), 제 2 배치채널의 2부분(B2), 제 3 배치채널의 2부분(C2), 및 제 4 배치채널의 2부분(D2)이 채워진다. 프로세싱 엘리먼트(PE)는 가중치(W)와 제 1, 제 2, 제 3 및 제 4 배치채널들의 2부분들 각각에 대한 계산을 수행한다.

[0258] S3에서 가중치(W)는 S1에서와 같이 유지된 상태에서 제 1 배치채널의 3부분(A3), 제 2 배치채널의 3부분(B3), 제 3 배치채널의 3부분(C3) 및 제 4 배치채널의 3부분(D3)이 채워진다. 프로세싱 엘리먼트(PE)는 가중치(W)와 제 1, 제 2, 제 3 및 제 4 배치채널들의 3부분들 각각에 대한 계산을 수행한다.

[0259] 이후 S4에서 가중치(W)는 S1에서와 같이 유지된 상태에서 제 1 배치채널의 4부분(A4), 제 2 배치채널의 4부분(B4), 제 3 배치채널의 4부분(C4) 및 제 4 배치채널의 4부분(D4)이 채워진다. 프로세싱 엘리먼트(PE)는 가중치(W)와 제 1, 제 2, 제 3 및 제 4 배치채널들의 4부분들 각각에 대한 계산을 수행한다.

[0260] 제 1, 제 2, 제 3, 및 제 4 배치채널들에 대한 계산이 완료되면 특징맵이 생성되고, 활성화 맵이 선택적으로 적용되어 활성화 맵이 생성될 수도 있다. 이와 같이 생성된 특징맵 또는 활성화 맵은 또 다른 컨볼루션 연산을 위해 컨볼루션 레이어로 입력되거나, 풀링 연산을 위해 풀링 레이어로 입력되거나, 분류를 위해 완전 연결 레이어로 입력될 수 있으나, 이에 한정되지 않는다. 이러한 계산들은 앞서 서술한 바와 같이 프로세싱 엘리먼트(PE)에 의해서 수행된다.

[0261] 이처럼, 프로세싱 엘리먼트(PE)는 메모리(즉, 내부 메모리)에 계속 유지되는 일 세트의 가중치와 복수의 배치채널들 각각을 계산한다.

[0262] 도 5 내지 도 6에서 제안된 동작 방식의 배치모드는 각각의 배치채널의 레이어별로 특징맵만 타일링(tiling)하

는 방식으로 설명될 수 있으며 제 1 배치모드로 지칭될 수 있다. 제 1 배치모드는 인공신경망모델의 레이어 중에서 특징맵의 파라미터 크기가 커널의 파라미터 크기보다 상대적으로 큰 경우에 활용될 수 있다.

[0263] 한편, 종래에는 복수의 연속하는 데이터 또는 연속하는 이미지 데이터들을 각각 처리할 때마다 가중치가 연산마다 새로 액세스되었다. 이러한 종래의 방식은 비효율 적이다.

[0264] 반면에, 본 개시에 따른 신경 프로세싱 유닛은 가중치를 메모리에 계속 유지함으로써, 가중치가 새로 액세스되는 것이 최소화되어 처리 속도가 향상되고 소비되는 에너지도 감소된다. 본 예시에서 메모리는 NPU 내부 메모리뿐만 아니라 온칩 메모리나 메인 메모리인 경우에도 동일하게 성능 향상과 에너지 감소 효과를 가진다.

[0265] 도 7은 본 개시의 다른 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다. 도 8은 본 개시의 다른 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

[0266] 제시된 예시에서 복수의 배치채널은 제 1 배치채널(Batch A), 및 제 2 배치채널(Batch B)을 포함한다. 각각의 배치채널은 예를 들어 4 개의 부분(또는 4 개의 일부)으로 나뉘어 질 수 있다.

[0267] 먼저, 도 7을 참조하면 일 세트의 가중치, 제 1 배치채널의 적어도 일부, 및 제 2 배치채널의 적어도 일부가 적어도 하나의 메모리에 저장된다(S2011). 여기서, 일 세트의 가중치는 적어도 하나의 가중치 값을 포함하는 가중치 행렬 중 적어도 하나의 가중치 값을 의미할 수 있다. 또한 메모리는 내부 메모리, 온칩 메모리 또는 메인 메모리일 수 있다.

[0268] 다음으로, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부 각각과 일 세트의 가중치가 계산된다(S2013). 해당 계산은 예를 들어 간격에 따른 컨볼루션 연산에 대응할 수 있다.

[0269] 다음으로, 일 세트의 가중치를 유지하면서 제 1 배치채널의 다른 일부 및 제 2 배치채널의 다른 일부가 적어도 하나의 메모리에 저장되고(S2015), 제 1 배치채널의 다른 일부 및 제 2 배치채널의 다른 일부 각각과 일 세트의 가중치가 계산된다(S2017).

[0270] 이후 또 다른 세트의 가중치가 적어도 하나의 메모리에 저장되고, 또 다른 세트의 가중치를 사용하여 인공신경망 연산이 수행된다(S2019). 여기서, 또 다른 세트의 가중치는 적어도 하나의 가중치 값을 포함하는 가중치 행렬 중 다른 하나의 가중치 값을 의미할 수 있다.

[0271] 이에 대해서 구체적으로 도 8을 참조하여 설명하면, 본 예시에서 메모리는 3개의 메모리 공간을 가진다고 가정한다.

[0272] 본 예시에서 일 세트의 가중치는 제 1 가중치(W1), 제 2 가중치(W2), 제 3 가중치(W3), 제 4 가중치(W4) 중 적어도 하나이고, 또 다른 세트의 가중치는 제 1 가중치(W1), 제 2 가중치(W2), 제 3 가중치(W3), 제 4 가중치(W4) 중 다른 하나인 것으로 가정한다.

[0273] 본 예시에서 제 1 배치채널(Batch A)은 A1, A2, A3, A4를 포함하고, 제 2 배치채널(Batch B)은 B1, B2, B3, B4를 포함한다고 가정한다. 또 다른 세트의 가중치는 다음 세트의 가중치로 지칭될 수 있다.

[0274] 도 8을 참조하면, S1에서 3개의 메모리 공간에는 일 세트의 가중치인 제 1 가중치(W1), 제 1 배치채널의 1부분(A1), 제 2 배치채널의 1부분(B1)이 채워진다. 프로세싱 엘리먼트(PE)는 제 1 가중치(W1)와 제 1 및 제 2 배치채널들의 1부분들(A1, B1) 각각에 대한 계산을 수행한다.

[0275] S2에서 제 1 가중치(W1)는 S1에서와 같이 유지된 상태에서 제1 배치채널의 2부분(A2), 제 2 배치채널의 2부분(B2)이 채워진다. 프로세싱 엘리먼트(PE)는 제 1 가중치(W1)와 제 1 및 제 2 배치채널들의 2부분들(A2, B2) 각각에 대한 계산을 수행한다.

[0276] 이러한 제 1 가중치(W1)는 제 1, 및 제 2 배치채널의 제 3 부분들(A3, B3) 및 제 4 부분들(A4, B4) 각각에 대한 계산이 수행되는 S3 내지 S4 동안 더 유지될 수 있다.

[0277] 제 1 가중치(W1)와 제 1 및 제 2 배치채널들 각각에 대한 계산이 완료되면 S5에서 3개의 메모리 공간에는 또 다른 세트의 가중치인 제 2 가중치(W2), 제 1 배치채널의 1부분(A1), 제 2 배치채널의 1부분(B1)이 채워진다. 프로세싱 엘리먼트(PE)는 제 2 가중치(W2)와 제 1 및 제 2 배치채널들의 1부분들 각각에 대한 계산을 수행한다.

[0278] 이러한 제 2 가중치(W2)는 제 1, 및 제 2 배치채널의 제 2 부분들(A2, B2), 제 3 부분들(A3, B3) 및 제 4 부분들(A4, B4) 각각에 대한 계산이 수행되는 S6 내지 S8 동안 유지될 수 있다.

[0279] 제 2 가중치(W2)와 제 1 및 제 2 배치채널들 각각에 대한 계산이 완료되면 S9에서 3개의 메모리 공간에는 또 다른 세트의 가중치인 제 3 가중치(W3), 제 1 배치채널의 1부분(A1), 제 2 배치채널의 1부분(B1)이 채워진다. 프로세싱 엘리먼트(PE)는 제 3 가중치(W3)와 제 1 및 제 2 배치채널들의 1부분들 각각에 대한 계산을 수행한다.

[0280] 이러한 제 3 가중치(W3)는 제 1, 및 제 2 배치채널의 제 1 부분들(A1, B1), 제 2 부분들(A2, B2), 제 3 부분들(A3, B3) 및 제 4 부분들(A4, B4) 각각에 대한 계산이 수행되는 S10 내지 S12 동안 유지될 수 있다.

[0281] 제 3 가중치(W3)와 제 1 및 제 2 배치채널들 각각에 대한 계산이 완료되면 S13에서 3개의 메모리 공간에는 또 다른 세트의 가중치인 제 4 가중치(W4), 제 1 배치채널의 1부분(A1), 제 2 배치채널의 1부분(B1)이 채워진다. 프로세싱 엘리먼트(PE)는 제 4 가중치(W4)와 제 1 및 제 2 배치채널들의 1부분들 각각에 대한 계산을 수행한다.

[0282] 이러한 제 4 가중치(W4)는 제 1, 및 제 2 배치채널의 제 1 부분들(A1, B1), 제 2 부분들(A2, B2), 제 3 부분들(A3, B3) 및 제 4 부분들(A4, B4) 각각에 대한 계산이 수행되는 S14 내지 S16 동안 유지될 수 있다.

[0283] 제 1, 및 제 2 배치채널들에 대한 계산이 완료되면 특징맵이 생성되고, 활성화 맵이 적용되어 활성화 맵이 생성된다. 이와 같이 생성된 활성화 맵은 또 다른 컨볼루션 연산을 위해 컨볼루션 레이어로 입력되거나, 풀링 연산을 위해 풀링 레이어로 입력되거나, 분류를 위해 완전 연결 레이어로 입력될 수 있으나, 이에 한정되지 않는다. 이러한 계산들은 앞서 서술한 바와 같이 프로세싱 엘리먼트(PE)에 의해서 수행된다.

[0284] 이처럼 프로세싱 엘리먼트(PE)는 메모리에 계속 유지되는 복수의 가중치 값들과 복수의 배치채널들 각각을 계산한다.

[0285] 도 7 내지 도 8에서 제안된 동작 방식의 배치모드는 각각의 배치채널의 레이어별로 가중치와 특징맵을 각각 타일링하는 방식으로 설명될 수 있으며 제 2 배치모드로 지칭될 수 있다. 제 2 배치모드는 인공신경망모델의 레이어 중에서 가중치의 파라미터 크기와 특징맵의 파라미터 크기가 메모리보다 상대적으로 큰 경우에 활용될 수 있다.

[0286] 한편, 종래에는 복수의 연속하는 데이터 또는 연속하는 이미지 데이터들을 처리할 때 복수의 가중치 값들이 연산마다 새로 액세스되었다. 이러한 종래의 방식은 비효율적이다.

[0287] 반면에, 본 개시에 따른 신경 프로세싱 유닛은 복수의 가중치 값들을 메모리에 계속 유지함으로써, 복수의 가중치 값들이 새로 액세스되는 것이 최소화되어 처리 속도가 향상되고 소비되는 에너지도 감소된다. 본 예시에서 메모리는 NPU 내부 메모리뿐만 아니라 온칩 메모리나 메인 메모리인 경우에도 동일하게 성능 향상과 에너지 감소 효과를 가진다.

[0288] 도 9는 본 개시의 또 다른 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다. 도 10은 본 개시의 또 다른 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

[0289] 제시된 예시에서 복수의 배치채널은 제 1 배치채널(Batch A), 제 2 배치채널(Batch B), 제 3 배치채널(Batch C), 및 제 4 배치채널(Batch D)를 포함한다. 일 세트의 가중치는 예를 들어 4 개의 부분(또는 4 개의 일부)으로 나뉘어 질 수 있다.

[0290] 먼저, 도 9를 참조하면, 일 세트의 가중치, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부가 적어도 하나의 메모리에 저장되고(S2021), 제 1 배치채널의 적어도 일부와 일 세트의 가중치가 계산된다(S2023). 여기서, 일 세트의 가중치는 적어도 하나의 가중치 값을 포함하는 가중치 행렬 중 적어도 하나의 가중치 값을 의미할 수 있다. 또한 메모리는 내부 메모리, 온칩 메모리 또는 메인 메모리일 수 있다.

[0291] 다음으로, 제 2 배치채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안, 제 1 배치채널의 적어도 일부의 공간에 다음에 계산될 제 3 배치채널의 적어도 일부가 저장된다(S2025). 즉, 계산과 동시에 다음 계산될 파라미터가 메모리에 로드된다.

[0292] 다음으로, 제 3 배치채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안, 제 2 배치채널의 적어도 일부의 공간에 다음에 계산될 제 4 배치채널이 저장되고(S2027), 제 4 배치채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안, 제 3 배치채널의 적어도 일부의 공간에 다음에 계산될 제 1 배치채널의 적어도 일부가 저장된다(S2029).

[0293] 이에 대해서 구체적으로 도 10을 참조하여 설명하면, 본 예시에서 메모리는 3개의 메모리 공간을 가진다고 가정한다. 또한 본 예시에서 일 세트의 가중치는 제 1 가중치(W1), 제 2 가중치(W2), 제 3 가중치(W3), 제 4 가중치

(W4) 중 적어도 하나인 것으로 가정한다.

[0294] 도 10을 참조하면, S1에서 3개의 메모리 공간에는 일 세트의 가중치인 제 1 가중치(W1), 제 1 배치채널(A), 제 2 배치채널(B)이 채워진다. 프로세싱 엘리먼트(PE)는 제 1 가중치(W1)와 제 1 배치채널(A)에 대한 계산을 수행한다. S2에서 프로세싱 엘리먼트(PE)가 제 1 가중치(W1)와 제 2 배치채널(B)에 대한 계산을 수행하는 동안, 제 1 배치채널(A)의 메모리 주소에 해당하는 공간에 제 3 배치채널(C)이 로드된다. 이처럼 배치채널에 대한 계산과 다음 계산될 파라미터에 대한 로드가 동시에 수행되므로, 신경 프로세싱 유닛의 계산 속도가 더욱 빨라진다.

[0295] 다음으로, S3에서 프로세싱 엘리먼트(PE)가 제 1 가중치(W1)와 제 3 배치채널(C)에 대한 계산을 수행하는 동안, 제 2 배치채널(B)의 메모리 주소에 해당하는 공간에 제 4 배치채널(D)이 로드된다. S4에서 프로세싱 엘리먼트(PE)가 제 1 가중치(W1)와 제 4 배치채널(D)에 대한 계산을 수행하는 동안, 제 3 배치채널(C)의 메모리 주소에 해당하는 공간에 제 1 배치채널(A)이 로드된다.

[0296] 제 1 가중치(W1)와 각 배치채널간의 계산이 완료되면 S5에서 제 1 가중치(W1)의 메모리 주소에 해당하는 공간에는 제 2 가중치(W2)가 로드된다. S5에서 프로세싱 엘리먼트(PE)가 일 세트의 가중치인 제 2 가중치(W2)와 제 1 배치채널(A)에 대한 계산을 수행하는 동안, 제 4 배치채널(D)의 메모리 주소에 해당하는 공간에 제 2 배치채널(B)이 로드된다. 다양한 예시에서 제 2 가중치(W2)는 제 1 가중치(W1)가 계산되는 동안 다른 메모리 주소에 해당하는 공간에 로드될 수도 있다.

[0297] 제 2 가중치(W2)와 각 배치채널간의 계산이 완료되면 일 세트의 가중치인 제 3 가중치(W3)와 각 배치채널 간의 계산이 수행될 수 있다. 이러한 계산은 전술한 계산과 유사하게 수행될 수 있다.

[0298] 제 3 가중치(W3)와 각 배치채널간의 계산이 완료되면 일 세트의 가중치인 제 4 가중치(W4)와 각 배치채널간의 계산이 수행되며, 이러한 계산 또한 전술한 계산과 유사하게 수행될 수 있다. 예를 들어, SN에서 프로세싱 엘리먼트(PE)가 제 4 가중치(W4)와 제 4 배치채널(D)간의 계산을 수행하는 동안, 제 3 배치채널(C)의 메모리 주소에 해당하는 공간에 제 1 가중치(W1)와 제 1 배치채널(A)과의 계산 값(즉, 연산 값)(A')이 로드된다.

[0299] 제 4 가중치(W4)와 각 배치채널간의 계산이 완료되면 SN+1에서 제 4 가중치(W4)의 메모리 주소에 해당하는 공간에는 다음 프로세싱을 위해 파라미터(X)가 로드된다. SN+1에서 프로세싱 엘리먼트(PE)가 파라미터(X)와 계산 값(A')에 대한 계산을 수행하는 동안 제 4 배치채널(D)의 메모리 주소에 해당하는 공간에 제 1 가중치(W1)와 제 2 배치채널(B)과의 계산 값(B')이 로드된 후 다음 연산이 수행될 수 있다.

[0300] 제 1, 제 2, 제 3, 및 제 4 배치채널들에 대한 계산이 완료되면 특징맵이 생성되고, 활성화 맵이 적용되어 활성화 맵이 생성된다. 이와 같이 생성된 활성화 맵은 또 다른 컨볼루션 연산을 위해 컨볼루션 레이어로 입력되거나, 풀링 연산을 위해 풀링 레이어로 입력되거나, 분류를 위해 완전 연결 레이어로 입력될 수 있으나, 이에 한정되지 않는다. 이러한 계산들은 앞서 서술한 바와 같이 프로세싱 엘리먼트(PE)에 의해서 수행된다.

[0301] 이처럼 프로세싱 엘리먼트(PE)는 메모리에 계속 유지되는 복수의 가중치와 복수의 배치채널들 각각을 계산한다.

[0302] 부연 설명하면, S1에서 S2로 넘어가는 단계처럼, 제 1 배치채널(A)의 적어도 일부를 제 3 배치채널(C)이 덮어쓰는 방식으로 처리될 수 있다. 즉, 특정 시간 동안에는 제 1 배치채널(A)이 저장된 메모리 공간을 제 3 배치채널(C)이 점진적으로 채워 나갈 수 있다. 이때 제 3 배치채널(C)이 덮어쓰는 메모리 공간은 W1 가중치를 이용하여 컨볼루션을 완료한 제 1 배치채널(A)의 데이터가 저장된 메모리 공간일 수 있다.

[0303] 즉, 연산을 끝낸 입력 특징맵이 저장된 메모리 공간에서는, 특정 배치채널이 저장된 메모리 공간을 다른 배치채널이 점진적으로 채워 나갈 수 있다.

[0304] 도 9 내지 도 10에서 제안된 동작 방식의 배치모드는, 복수의 배치채널(예를 들어 4개의 채널) 중 적어도 일부의 복수의 배치채널(예를 들어 2개의 채널)의 파라미터가 메모리에 각각 저장되고, 하나의 배치채널의 파라미터 연산이 완료되고, 다른 배치채널의 파라미터를 계산할 때, 다음에 계산할 다른 배치채널의 파라미터를 상기 파라미터 연산이 완료된 상기 하나의 배치채널의 메모리 영역에 로드되는 방식으로 설명될 수 있으며, 제 3 배치모드로 지칭될 수 있다. 제 3 배치모드는 전체 배치채널의 개수만큼 메모리 영역을 분할하지 않고, 전체 배치채널의 개수보다 적은 개수로 메모리 영역을 분할하기 때문에, 메모리에 할당된 각각의 영역의 크기를 키울 수 있다.

[0305] 한편, 종래에는 복수의 연속하는 데이터 또는 연속하는 이미지 데이터들을 처리할 때 가중치가 연산마다 새로 액세스되었다. 이러한 종래의 방식은 비효율적이다.

[0306] 반면에, 본 개시에 따른 신경 프로세싱 유닛은 가중치와 배치채널이 계산되는 동안 사용이 완료된 메모리 주소에 해당하는 공간에 새로운 배치채널 또는 새로운 가중치를 로드하여 사용하지 않는 메모리 공간을 최대한 활용함으로써, 처리 속도가 향상되고 소비되는 에너지도 감소된다. 본 예시에서 메모리는 NPU 내부 메모리뿐만 아니라 온칩 메모리나 메인 메모리인 경우에도 동일하게 성능 향상과 에너지 감소 효과를 가진다.

[0307] 도 11은 본 개시의 다양한 예시에 따른 신경 프로세싱 유닛이 동작하는 방법을 설명하는 예시적인 순서도이다. 도 12는 본 개시의 다양한 예시에 따른 신경 프로세싱 유닛에서 인공신경망 파라미터들이 할당되는 메모리 공간을 단계에 따라 나타낸 예시적인 개략도이다.

[0308] 제시된 예시에서 복수의 배치채널은 제 1 배치채널(Batch A), 제 2 배치채널(Batch B), 제 3 배치채널(Batch C), 및 제 4 배치채널(Batch D)를 포함한다. 일 세트의 가중치는 예를 들어 2 개의 부분으로 나뉘어 질 수 있다.

[0309] 먼저, 도 11을 참조하면 일 세트의 가중치, 제 1 배치채널의 적어도 일부, 및 제 2 배치채널의 적어도 일부가 적어도 하나의 메모리에 저장되고(S2031), 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부 각각과 일 세트의 가중치가 계산되고, 계산된 값이 적어도 하나의 메모리에 저장된다(S2033). 여기서, 일 세트의 가중치는 적어도 하나의 가중치 값을 포함하는 가중치 행렬을 의미하고, 메모리는 내부 메모리, 온칩 메모리 또는 메인 메모리일 수 있다. 본 예시에서 계산된 값은 일 세트의 가중치, 제 1 배치채널의 적어도 일부, 및 제 2 배치채널의 적어도 일부와 동일한 메모리에 저장될 수 있다.

[0310] 다음으로, 다음 프로세싱 단계를 위해 또 다른 세트의 가중치가 적어도 하나의 메모리에 저장되고(S2035), 계산된 값들과 다음 프로세싱 단계를 위한 또 다른 세트의 가중치가 계산된다(S2037). 해당 계산은 ReLU 연산 또는 다음 스테이지 컨볼루션 연산에 대응할 수 있으나, 이에 한정되지 않는다. 이를 통해 계산된 값들은 적어도 하나의 메모리에 저장되고, 계산된 값들을 이용하여 인공신경망 연산이 수행된다(S2039).

[0311] 이와 같이 계산된 값들이 NPU 내부 메모리(200)에 유지되므로, 계산을 위해 메인 메모리 또는 외부 메모리를 이용할 필요가 없다.

[0312] 이에 대해서 구체적으로 도 12를 참조하여 설명하면, 본 예시에서 메모리는 10개의 메모리 공간을 가진다고 가정한다. 또한 본 예시에서 일 세트의 가중치는 제 1 가중치(W5)를 의미하고, 또 다른 세트의 가중치는 제 2 가중치(W6)를 의미한다고 가정한다.

[0313] 도 12를 참조하면, S1에서 5개의 메모리 공간에는 일 세트의 가중치인 제 1 가중치(W5), 제 1 배치채널(A), 제 2 배치채널(B), 제 3 배치채널(C), 및 제 4 배치채널(D)이 채워진다. 프로세싱 엘리먼트(PE)는 제 1 가중치(W5)와 제 1, 제 2, 제 3, 및 제 4 배치채널들 각각에 대한 계산을 수행한다. 계산된 값들(A', B', C', D')은 4개의 메모리 공간에 저장되고, 다음 프로세싱 단계를 위해 또 다른 세트의 가중치인 제 2 가중치(W6)가 하나의 메모리 공간에 로드된다.

[0314] 다음으로, S2에서 제 2 가중치(W6)는 S1에서와 같이 유지된 상태에서 제 2 가중치(W6)와 계산된 값들(A', B', C', D') 각각에 대한 계산이 수행된다. 이러한 계산은 프로세싱 엘리먼트(PE)에 의해 수행되며, 예를 들어 ReLU 연산 또는 다음 스테이지 컨볼루션 연산에 대응할 수 있다.

[0315] 사용된 제 1 가중치(W5)는 메모리에서 삭제된다. 제 2 가중치(W6)와 계산된 값들(A', B', C', D')과의 새롭게 계산된 값들(A'', B'', C'', D'')은 제 1 배치채널(A), 제 2 배치채널(B), 제 3 배치채널(C), 및 제 4 배치채널(D) 각각의 메모리 공간에 채워진다. 다음 파라미터 예를 들어 파라미터(X)가 하나의 메모리 공간에 로드된다.

[0316] 다양한 예시에서 계산된 값들이 적어도 하나의 메모리에 저장되는 동작 및 다음 프로세싱 단계를 위한 또 다른 세트의 가중치 또는 파라미터가 계산되는 동작은 내부 메모리에서만 이루어질 수 있다.

[0317] 이처럼, 계산된 값들은 메모리에 계속 유지될 수 있다.

[0318] 도 11 내지 도 12에서 제안된 동작 방식의 배치모드는 출력 특징맵이 다음 연산의 입력 특징맵으로 활용되는 특성을 활용하여 각각의 배치채널을 처리하는 방식으로 설명될 수 있으며 제 4 배치모드로 지칭될 수 있다.

[0319] 한편, 종래에는 복수의 연속하는 데이터 또는 연속하는 이미지 데이터들을 처리할 때 이미지 데이터들에 대한 계산 값은 메인 메모리에 저장되고, 연산 때마다 다음 계산을 위해 새로 액세스되었다. 이러한 종래의 방식은 비효율 적이다.

[0320] 반면에, 본 개시에 따른 신경 프로세싱 유닛은 계산된 값들을 NPU 내부 메모리(200)에 계속 유지함으로써, 계산된 값들이 새로 액세스되는 것이 최소화되어 처리 속도가 향상되고 소비되는 에너지도 감소된다. 본 예시에서 메모리는 NPU 내부 메모리(200)로 설명하였으나, 이에 한정되지 않으며, 온칩 메모리인 경우에 더 높아진 성능 향상과 에너지 감소 효과를 가진다.

[0321] 인공신경망모델은 복수의 레이어를 포함하고, 각각의 레이어는 가중치 파라미터, 특징맵 파라미터 정보를 포함한다. NPU 스케줄러는 상기 파라미터 정보를 제공받을 수 있다.

[0322] 본 개시에 따르면, 상술한 제 1 배치모드 내지 제 4 배치모드 중 적어도 하나의 배치모드를 선택적으로 활용하여 신경 프로세싱 유닛이 인공신경망모델을 처리하도록 구성될 수 있다.

[0323] 신경 프로세싱 유닛은 인공신경망모델의 각각의 레이어의 가중치 파라미터의 크기, 배치채널들의 특징맵 파라미터의 크기, 및 신경 프로세싱 유닛의 메모리 크기를 분석하여, 적어도 하나의 배치모드를 적용할 수 있다.

[0324] 신경 프로세싱 유닛은 인공신경망모델의 특정 그룹의 레이어들에 특정 배치모드를 적용할 수 있다.

[0325] 신경 프로세싱 유닛은 인공신경망모델의 하나의 레이어의 일부를 특정 배치모드로 처리되고 또 다른 일부를 다른 배치모드로 처리할 수 있다.

[0326] 신경 프로세싱 유닛은 인공신경망모델의 각각의 레이어마다 특정 배치모드를 적용할 수 있다.

[0327] 신경 프로세싱 유닛은 인공신경망모델의 하나의 레이어에 복수의 배치모드를 적용할 수 있다.

[0328] 신경 프로세싱 유닛은 인공신경망모델의 각각의 레이어 별로 최적의 배치모드를 제공하도록 구성될 수 있다.

[0329] 도 13은 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템을 나타낸 예시도이다.

[0330] 도 13을 참조하면, 자율 주행 시스템(C)은 자율 주행을 위한 복수의 센서를 구비하는 자율 주행 차량 및 복수의 센서로부터 획득된 센싱 데이터에 기반하여 차량이 자율 주행을 수행하도록 차량을 제어하는 차량 제어 장치(10000)를 포함할 수 있다.

[0331] 자율 주행 차량은 복수의 센서를 구비하고, 복수의 센서를 통해 차량 주변을 모니터링하여 자율 주행을 수행할 수 있다.

[0332] 자율 주행 차량에 구비된 복수의 센서는 자율 주행을 위해 요구될 수 있는 다양한 센서를 포함할 수 있다. 예를 들어, 다양한 센서는 이미지 센서, 레이더(Lidar), 및/또는 라이더(Radar), 및/또는 초음파(Ultrasound) 센서 등을 포함할 수 있다. 또한, 복수의 센서는 동일 센서가 복수로 구비되거나, 서로 다른 센서가 복수로 구비될 수 있다.

[0333] 이미지 센서는 전방 카메라(410), 좌측 카메라(430), 우측 카메라(420) 및 후방 카메라(440)에 대응할 수 있다. 다양한 예시에서 이미지 센서는 360도 카메라 또는 서라운드 뷰(surround view) 카메라에 대응할 수 있다.

[0334] 이미지 센서는 CMOS(Complementary Metal Oxide Semiconductor) 센서 및 CCD(Charge Coupled Device) 센서 등과 같이 컬러 영상(예: RGB(380nm~680nm) 영상)을 촬영하기 위한 이미지 센서를 포함할 수 있으나, 이에 한정되지 않는다.

[0335] 다양한 예시에서 이미지 센서는 자율 주행 차량의 주간 환경뿐만 아니라 야간 환경을 촬영하기 위한 IR(Infrared Ray) 센서, 및/또는 NIR(Near IR) 센서 등을 더 포함할 수 있으나, 이에 한정되지 않는다. 이러한 센서들은 컬러 이미지 센서에 의해 야간 환경에서 촬영된 야간 저조도 영상의 품질을 보상하기 위해 이용될 수 있다. 여기서, NIR 센서는 CMOS 센서의 RGB 및 IR 센서의 구조를 결합한 쿼드 화소(Quad pixel) 형태로 구현될 수 있으나, 이에 한정되지 않는다.

[0336] NIR 센서를 이용하여 근적외선 영상을 촬영하기 위해 자율 주행 차량은 NIR 광원(예: 850nm~940nm)을 더 구비할 수 있다. 이러한 NIR 광원은 사람의 시각에 감지되지 않아 다른 운전자의 시야에 방해되지 않으며, 차량 전조등에 추가 조명으로 제공될 수 있다.

[0337] 다양한 예시에서 IR 센서는 열감지 센서 로서, 열화상 영상을 촬영하기 위해 이용될 수 있다. 다양한 예시에서 자율 주행 차량은 IR 센서에 대응하여 IR 광원을 더 구비할 수 있다.

[0338] 예를 들어, 열화상 영상은 RGB 영상과 이에 동기화된 열감지 정보를 포함하도록 구성될 수 있다. 또한, 열화상 영상은 자율 주행 시 위험 요소에 해당할 수 있는 도로 표면 온도, 차량의 엔진, 배기구, 야간 환경에서의 야생

동물, 및/또는 빙판길 등을 인식하기 위해 이용될 수 있다.

[0339] 다양한 예시에서 IR 센서는 자율 주행 차량의 실내에 구비될 경우 열감지를 통해 운전자(또는 사용자)의 온도를 감지하여 운전자의 고열, 감기, 코로나 감염 여부, 및/또는 실내 공조 등의 상태를 결정하기 위해 이용될 수 있다.

[0340] 이를 통해서 촬영된 열화상 영상은 객체 인식을 위해 후술할 인공신경망모델의 학습을 위한 참조 영상으로서 이용될 수 있다.

[0341] 다양한 예시에서 IR 광원은 복수의 IR 센서와 동기화될 수 있으며, 이와 같이 동기화된 IR 광원 및 복수의 IR 센서에 의해서 촬영된 열화상 영상은 객체 인식을 위해 후술할 인공신경망모델의 학습을 위한 참조 영상으로서 이용될 수 있다.

[0342] 다양한 예시에서 IR 광원은 전방의 조사각을 가지고, 이러한 조사각은 차량의 헤드라이트와 다른 조사각일 수 있다.

[0343] 다양한 예시에서 NIR 광원 및 IR 광원은 프레임마다 온/오프되어 재귀반사(retro-reflector) 특성을 갖는 객체(예: 안전띠, 표지판, 및 스텔스 차량 전조등 등)를 인식하기 위해 이용될 수 있다.

[0344] 이와 같이 이미지 센서를 포함하는 복수의 카메라가 자율 주행 차량의 다양한 위치에, 다양한 개수만큼 구비될 수 있다. 여기서, 다양한 위치 및 다양한 개수는 자율 주행을 위해 요구되는 위치 및 개수일 수 있다.

[0345] 복수의 카메라(410, 420, 430, 440)는 차량 주변(예: 실외)의 영상을 촬영하고, 촬영된 복수의 영상을 차량 제어 장치로 전달할 수 있다. 복수의 영상은 동일 시간에 촬영된 컬러 영상(예: RGB 영상)과 함께 적외선 영상 및 근적외선 영상(또는 열화상 영상) 등 중 적어도 하나를 포함하거나, 둘 이상의 조합에 의해서 형성된 영상을 포함할 수 있으나, 이에 한정되지 않는다.

[0346] 다양한 예시에서 복수의 카메라(410, 420, 430, 440)는 자율 주행 차량의 실내에 구비될 수도 있다. 이와 같이 실내에 구비된 복수의 카메라는 다양한 위치에 배치되고, 이를 통해서 촬영된 영상은 운전자 상태 감지 시스템(Driven State Monitoring)을 위해 이용될 수 있으나, 이에 한정되지 않는다. 다양한 예시에서 촬영된 영상은 운전자의 졸음, 음주, 영유아 방치, 편의 및 안전 등과 같은 상태를 결정하기 위해 이용될 수 있다.

[0347] 자율 주행 차량은 차량 제어 장치(10000)로부터 차량 주행 지시를 수신하고, 수신된 차량 주행 지시에 따라 차량 주행을 수행할 수 있다.

[0348] 다음으로, 차량 제어 장치(10000)는 복수의 센서로부터 획득된 센싱 데이터에 기반으로 하여 자율 주행 차량을 제어하기 위한 전자 장치일 수 있다. 이러한 차량 제어 장치(10000)는 차량에 탑재 가능한 전장 시스템으로 구현되거나, 블랙박스 등과 같은 차량에 탈부착 가능한 장치로 구현되거나, 스마트폰, PDA(Personal Digital Assistant), 및/또는 태블릿 PC(Personal Computer) 등과 같은 휴대 장치로 구현될 수 있으나, 이에 한정되지 않는다. 차량 제어 장치(10000)는 프로세서를 포함할 수 있다. 프로세서는 중앙 처리 장치(CPU), 그래픽 처리 장치(GPU), 어플리케이션 프로세서(AP), 디지털 신호 처리 장치(DSP), 산술 논리 연산 장치(ALU) 및 인공신경망 프로세서(NPU) 중 적어도 하나를 포함하도록 구성될 수 있다. 단, 본 개시의 프로세서는 상술한 프로세서들에 제한되지 않는다.

[0349] 구체적으로, 차량 제어 장치(10000)는 복수의 센서로부터 획득된 센싱 데이터를 이용하여 자율 주행을 위해 사용되는 자율 주행 지도 데이터를 생성하고, 생성된 자율 주행 지도 데이터를 기반으로 하는 자율 주행 지시를 자율 주행 차량으로 전달할 수 있다.

[0350] 여기서, 자율 주행 지도 데이터는 카메라, 레이더, 라이더 및/또는 초음파 센서 등의 다양한 센서에 의해서 측정된 센싱 데이터를 기반으로 감지된 차량의 주변 환경을 정밀하게(또는 상세하게) 나타낸 지도 데이터로서, 3차원으로 구현될 수 있다.

[0351] 이러한 자율 주행 지도 데이터를 생성하기 위해 차량 제어 장치(10000)는 센싱 데이터를 기반으로 도로 환경 데이터 및 실시간 환경 데이터를 검출할 수 있다. 도로 환경 데이터는 차선, 가드레일, 도로 곡률/경사, 신호등/표지판 위치, 및/또는 교통 표식 등을 포함할 수 있으나, 이에 한정되지 않는다. 실시간 환경 데이터는 시시각각 변화하는 주변 환경 데이터로서, 전후방 접근 차량, 공사(또는 사고) 구간, 차량 흐름, 실시간 신호 정보, 노면 상태, 장애물, 및/또는 보행자 등을 포함할 수 있으나, 이에 한정되지 않는다. 다양한 예시에서 도로 환경 데이터 및 실시간 환경 데이터는 지속적으로 업데이트될 수 있다.

[0352] 자율 주행 지도 데이터는 상술한 바와 같이 센싱 데이터를 이용하여 생성할 수 있으나, 이에 한정되지 않으며, 특정 지역에 대해서 기 생성된 지도 데이터를 이용할 수도 있다.

[0353] 기 생성된 지도 데이터는 다양한 센서가 구비된 측량 차량에 의해서 미리 수집된 도로 환경 데이터 및 주변 환경 데이터 중 적어도 일부를 포함하며, 클라우드 기반 데이터베이스에 저장될 수 있다. 이러한 지도 데이터는 실시간으로 분석되어 지속적으로 업데이트될 수 있다.

[0354] 차량 제어 장치(10000)는 자율 주행 차량의 위치에 해당하는 지역의 지도 데이터를 데이터베이스로부터 획득하고, 획득된 지도 데이터와 함께 다양한 센서에 의해서 측정된 센싱 데이터를 기반으로 자율 주행 지도 데이터를 생성할 수 있다. 예를 들어, 차량 제어 장치(10000)는 센싱 데이터를 기반으로 특정 지역에 관련하여 획득된 지도 데이터를 실시간으로 업데이트하여 자율 주행 지도 데이터를 생성할 수 있다.

[0355] 차량 제어 장치(10000)는 차량의 자율 주행을 제어하기 위해 실시간으로 급격하게 변화하는 주변 환경을 정확하게 인식할 필요가 있다. 다시 말해서, 차량의 자율 주행 시 위험 상황을 미리 대처하도록 차량 주변에서 고려될 수 있는 목표 객체(또는 객체)를 정확하고, 지속적으로 인식할 필요가 있다. 여기서, 목표 객체는 전후방 접근 차량, 신호등, 신호등의 실시간 신호 정보, 장애물, 통행인, 사람, 동물, 도로, 표지판, 도로선 및 보행자 등 중 적어도 하나를 포함할 수 있으나, 이에 한정되지 않는다.

[0356] 차량의 주변 환경에서 목표 객체에 대한 정확하고, 지속적인 인식이 이루어지지 않을 경우 차량과 목표 객체 사이에 큰 충돌이 발생하는 등의 위험 상황이 발생되어 차량의 자율 주행이 안전하고, 올바르게 이루어지지 않을 수 있다.

[0357] 차량의 안전한 자율 주행을 위해 차량 제어 장치(10000)는 복수의 카메라(410, 420, 430, 440)로부터 수신된 복수의 영상에 기반하여 자율 주행에 관련된 목표 객체를 인식할 수 있다. 여기서, 복수의 영상은 복수의 카메라(410, 420, 430, 440)를 통해 동일 시간에 촬영된 영상일 수 있다.

[0358] 목표 객체에 관한 정확하고, 지속적인 인식을 위해 차량의 다양한 주변 환경에 관한 학습 데이터를 기반으로 목표 객체를 인식하도록 학습된 인공지능 기반 객체 인식 모델(또는 인공신경망모델)(즉, ANN)이 이용될 수 있다. 여기서, 학습 데이터는 차량의 다양한 주변 환경을 촬영한 복수의 참조 영상일 수 있으나, 이에 한정되지 않는다. 여기서, 복수의 참조 영상은 컬러 영상과 함께 적외선, 근적외선 및 열화상 영상 중 적어도 둘 이상을 포함할 수 있으나, 이에 한정되지 않는다. 다양한 예시에서 복수의 참조 영상은 이미지 센서(예: 컬러 이미지 센서, IR 센서 및/또는 NIR 센서), 라이더, 레이더 및 초음파 센서 중 적어도 둘 이상의 조합에 의해서 형성될 수도 있다.

[0359] 차량 제어 장치(10000)는 객체 인식 모델을 이용하여 복수의 카메라(410, 420, 430, 440) 로부터 수신된 영상에서 목표 객체를 인식하고, 인식된 목표 객체를 기반으로 하는 자율 주행 지시를 자율 주행 차량으로 전달할 수 있다. 예를 들어, 자율 주행 동안 도로 상에서 보행자와 같은 목표 객체가 인식되면 차량 제어 장치(10000)는 자율 주행 차량의 주행을 중지하기 위한 지시를 자율 주행 차량으로 전달할 수 있다.

[0360] 이와 같이 본 발명은 인공지능 기반 객체 인식 모델을 이용하여 자율 주행 차량의 자율 주행을 위해 고려될 수 있는 목표 객체를 인식함으로써, 정확하게 빠른 객체 인식이 가능하다.

[0361] 하기에서는 도 14를 참조하여 자율 주행 차량에 대해서 보다 구체적으로 설명하도록 한다.

[0362] 도 14는 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템의 개략적인 블록도이다.

[0363] 도 14를 참조하면, 자율 주행 차량은 통신부(600), 센서(400), 저장부 및 제어부를 포함한다. 제시된 예시에서 자율 주행 차량은 도 13의 자율 주행 시스템을 의미할 수 있다.

[0364] 통신부(600)는 자율 주행 차량이 외부 장치와 통신이 가능하도록 연결한다. 통신부(600)는 유/무선 통신을 이용하여 차량 제어 장치(10000)와 연결되어 자율 주행에 관련된 다양한 데이터를 송수신할 수 있다. 구체적으로, 통신부(600)는 복수의 센서로부터 획득된 센싱 데이터를 차량 제어 장치(10000)로 전달하고, 차량 제어 장치(10000)로부터 자율 주행 지시를 수신할 수 있다.

[0365] 위치 탐색부(700)는 자율 주행 차량의 위치를 탐색할 수 있다. 위치 탐색부(700)는 위성 항법 및 추측 항법 중 적어도 하나를 이용할 수 있다. 예를 들어, 위성 항법을 이용하는 경우 위치 탐색부(700)는 차량의 위치 정보를 측정하는 GPS(Global Positioning System), GLONASS(Global Navigation Satellite System), 갈릴레오(Galileo), 베이더우(Beidou) 등의 위치 탐색 시스템으로부터 위치 정보를 획득할 수 있다.

[0366] 다양한 예시에서 추측 항법을 이용하는 경우 위치 탐색부(700)는 차량의 속도계, 자이로센서 및 지자기 센서 등과 같은 움직임 센서(미도시)로부터 차량의 침로 및 속도 등을 계산하고, 이를 바탕으로 차량의 위치 정보를 추측할 수 있다.

[0367] 다양한 예시에서 위치 탐색부(700)는 위성 항법 및 추측 항법 둘 다를 이용하여 차량의 위치 정보를 획득할 수도 있다.

[0368] 센서(400)는 차량의 주변 환경을 감지하기 위해 사용되는 센싱 데이터를 획득할 수 있다. 센서(400)는 이미지 센서(410), 라이더(450), 레이더(460) 및 초음파 센서(470)를 포함할 수 있다. 동일 센서가 복수로 구비되거나, 서로 다른 센서가 복수로 구비될 수 있다.

[0369] 이미지 센서(410)는 차량의 주변을 촬영하기 위해 구비되며, CCD 센서, CMOS 센서, IR 센서 및/또는 NIR 센서 등 중 적어도 하나일 수 있다. 이러한 이미지 센서(410)는 복수로 구비될 수 있으며, 복수의 이미지 센서에 대응하여 복수의 카메라가 자율 주행 차량에 다양한 위치에 구비될 수 있다. 예를 들어, 차량의 주변을 촬영하기 위해 복수의 전방, 좌/우측방 및 후방 카메라가 구비되거나, 360도 카메라 또는 서라운드 뷰 카메라가 구비될 수 있으나, 이에 한정되지 않는다.

[0370] 다양한 예시에서 CCD 센서 및/또는 CMOS 센서에 대응하는 카메라는 차량 주변에 관한 컬러 영상을 획득할 수 있다.

[0371] 다양한 예시에서 IR 센서 및/또는 NIR 센서는 적외선 및/또는 근적외선을 기반으로 온도 등을 감지하여 객체를 검출할 수 있다. IR 센서 및/또는 NIR 센서에 대응하여 적외선 카메라, 근적외선 카메라 및/또는 열화상 카메라가 자율 주행 차량의 적어도 하나의 위치에 구비되며, 차량의 주변에 관한 적외선 영상, 근적외선 영상 및/또는 열화상 영상을 획득할 수 있다. 이와 같이 획득된 적외선 영상, 근적외선 영상 및/또는 열화상 영상은 빛이 취약한 장소(또는 어두운 장소)에서의 자율 주행을 위해 이용될 수 있다.

[0372] 라이더(450)는 전자기파를 발사하고, 주변 객체에서 반사되어 돌아오는 반향파를 이용하여 객체의 위치, 객체의 속도 및/또는 객체의 방향을 검출할 수 있다. 다시 말해서, 라이더(450)는 차량이 위치한 환경 내에서 객체를 감지하기 위한 센서일 수 있다.

[0373] 레이더(460)는 레이저를 발사하고, 주변 객체에서 반사되어 돌아오는 반사광을 이용하여 객체의 형상, 및/또는 객체와의 거리 등과 같은 주변 환경을 감지할 수 있는 센서일 수 있다.

[0374] 초음파 센서(470)는 초음파를 발사하고, 주변 객체에서 반사되어 돌아오는 초음파를 이용하여 차량과 객체 사이의 거리를 감지할 수 있다. 이러한 초음파 센서(470)는 차량과 객체 사이의 근거리 측정을 위해 이용될 수 있다. 예를 들어, 초음파 센서(470)는 차량의 전면 왼쪽, 전면 오른쪽, 왼쪽 측면, 후면 왼쪽, 후면 오른쪽 및 오른쪽 측면에 구비될 수 있으나, 이에 한정되지 않는다.

[0375] 저장부는 자율 주행을 위해 사용되는 다양한 데이터를 저장할 수 있다. 다양한 예시에서 저장부는 플래시 메모리 타입(flash memory type), 하드디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(Random Access Memory, RAM), SRAM(Static Random Access Memory), 롬(Read-Only Memory, ROM), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory), 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다.

[0376] 제어부는 통신부(600), 위치 탐색부(700), 센서(400), 및 저장부와 동작 가능하게 연결되며, 자율 주행을 위한 다양한 명령들을 수행할 수 있다. 제어부는 중앙 처리 장치(CPU), 그래픽 처리 장치(GPU), 어플리케이션 프로세서(AP), 디지털 신호 처리 장치(DSP), 및 산술 논리 연산 장치(ALU) 중 하나 그리고 인공신경망 프로세서(NPU)를 포함하도록 구성될 수 있다.

[0377] 구체적으로, 제어부는 센서를 통해서 획득된 센싱 데이터를 통신부(600)를 통해 차량 제어 장치(10000)로 전달한다. 여기서, 센싱 데이터는 자율 주행 지도 데이터를 생성하거나, 목표 객체를 인식하기 위해 이용될 수 있다. 예를 들어, 센싱 데이터는 이미지 센서(410)를 통해서 획득된 영상 데이터, 라이더(450)를 통해서 획득된 객체의 위치, 객체의 속도 및/또는 객체의 방향 등을 나타내는 데이터, 레이더(460)를 통해서 획득된 객체의 형상, 및/또는 객체와의 거리 등을 나타내는 데이터, 및/또는 초음파 센서(470)를 통해서 획득된 차량과 객체 사이의 거리를 나타내는 데이터 등을 포함할 수 있으나, 이에 한정되지 않는다. 여기서, 영상 데이터는 동일 시간에 촬영된 컬러 영상, 적외선 영상, 근적외선 영상 및 열화상 영상 중 복수의 영상이 포함될 수 있다. 다양한

예시에서 영상 데이터는 이미지 센서(410), 라이더(450), 레이더(460) 및 초음파 센서(470) 중 적어도 둘 이상의 조합에 의해서 형성될 수도 있다.

[0378] 제어부는 차량 제어 장치(10000)로부터 자율 주행 지시를 수신하고, 수신된 자율 주행 지시에 따라 차량의 자율 주행을 수행할 수 있다.

[0379] 도 15는 본 발명의 일 예시에 따른 신경 프로세싱 유닛이 탑재된 자율 주행 시스템에서 자율 주행을 위해 목표 객체를 인식하기 위한 발명을 설명하기 위한 순서도이다.

[0380] 도 15를 참조하면, 차량 제어 장치는 자율 주행 차량에 구비된 복수의 카메라로부터 복수의 영상을 수신한다(S1200). 여기서, 복수의 영상은 동일 시간에 촬영된 영상으로, 컬러 영상, 적외선 영상 및/또는 열화상 영상 등을 포함할 수 있다. 다시 말해서, 복수의 영상(또는 이미지)은 실질적으로 동일한 기간에 캡쳐된 이미지들을 의미한다. 이와 같이 컬러 영상 및 적외선 영상, 또는 컬러 영상 및 열화상 영상을 이용하는 경우 차량의 주간 자율 주행뿐만 아니라 야간 자율 주행 또한 원활하게 수행될 수 있다. 다시 말해서, 복수의 영상은 RGB 이미지, IR 이미지, RADAR 이미지, ULTRASOUND 이미지, LIDAR 이미지, 열 화상 이미지, 및 NIR 이미지 중 적어도 하나를 포함한다.

[0381] 차량 제어 장치는 복수의 영상을 연속적으로 배열한 배치 데이터를 생성할 수 있다(S1210). 여기서, 배치 데이터는 객체 인식 모델의 입력층을 구성하는 입력 노드에 대응될 수 있으며, 복수의 배치채널들을 의미할 수 있다. 이러한 복수의 배치채널들 각각은 복수의 이미지들 각각에 대응할 수 있다. 차량 제어 장치는 배치 데이터를 입력으로 자율 주행에 관련된 목표 객체를 인식하도록 학습된 객체 인식 모델을 이용하여 복수의 영상으로부터 목표 객체를 인식한다(S1220). 여기서, 객체 인식 모델은 차량의 다양한 주변 환경에 관련된 복수의 학습 영상을 입력으로 하여 복수의 학습 영상으로부터 목표 객체를 인식하도록 학습된 인공신경망모델을 의미한다. 차량의 다양한 주변 환경은 주간 및/또는 야간 환경을 포함하고, 이러한 환경에서 목표 객체를 정확하게 인식하기 위해 주간 환경을 촬영한 영상 및/또는 야간 환경을 촬영한 영상이 학습 영상으로 이용될 수 있다.

[0382] 본 예시에서 ANN(즉, 인공신경망모델)은 복수의 배치채널들로부터 객체 검출, 분류 또는 세그먼트화를 포함하는 적어도 하나의 동작을 수행한다. 이러한 ANN은 객체에 대한 향상된 검출을 위해 복수의 배치채널들을 전처리할 수 있고, 복수의 배치채널들을 전처리하면서 복수의 배치채널들로부터 객체를 동시에 검출하도록 구성될 수 있다. 여기서, 복수의 배치채널들은 RGB, YCBCR, HSV, 또는 HIS 컬러 공간들 중 어느 하나로부터의 채널들 및 IR 채널에 대응할 수 있다. 또한 복수의 배치채널들 각각은 차량의 내부를 캡처하는 이미지를 포함하고, ANN은 차량의 안전 관련 객체, 기능, 운전자의 상태, 및 승객의 상태 중 적어도 하나를 검출하도록 구성될 수 있다.

[0383] 다양한 예시에서 복수의 배치채널들 각각은 복수의 센서 데이터 각각에 대응하고, 복수의 센서 데이터는 압력 센서, 피에조 센서, 습도 센서, 먼지 센서, 스모그 센서, 소나(Sonar) 센서, 진동 센서, 가속도 센서 또는 모션 센서 중 하나 이상으로부터의 데이터를 포함할 수 있다. 이와 같이 목표 객체가 인식되면 차량 제어 장치는 인식된 목표 객체에 관련된 자율 주행 지시를 자율 주행 차량으로 전달하여 자율 주행 차량이 안전하게 자율 주행을 수행할 수 있도록 한다.

[0384] 본 개시에 따른 방법은 ANN (artificial neural network) 에 대한 동작들을 수행하는 단계를 포함하고, 동작들을 위해, 복수의 배치채널들은 제 1 배치채널 및 제 2 배치채널을 포함하고, 동작들은, 일 세트의 가중치, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부를 적어도 하나의 메모리에 저장하는 단계; 그리고 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치 값들을 계산하는 단계를 포함할 수 있다.

[0385] 제 1 배치채널의 적어도 일부의 크기 및 적어도 제 2 배치채널의 크기는 실질적으로 동일할 수 있다.

[0386] 상기 동작들을 위해, 일 세트의 가중치 값들은 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응할 수 있다.

[0387] 상기 동작들을 위해, 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 방법은, 제 2 배치채널의 적어도 일부와 일 세트 된 가중치 값들을 계산하는 동안, 적어도 하나의 메모리의 적어도 일부에 다음에 계산될 제 1 배치채널의 다른 적어도 일부를 저장하는 단계를 더 포함할 수 있다.

[0388] 상기 동작들을 위해, 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 더 포함하고, 동작들은, 일 세트의 가중치 값들을 유지하면서 적어도 하나의 메모리에 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부를 메모리에 저장하는 단계 및 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부와 일 세트의

가중치 값들을 계산하는 단계를 더 포함할 수 있다.

[0389] 상기 동작들을 위해, 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 그리고 일 세트의 가중치들은 대응되는 복수의 배치채널들 각각의 적어도 일부가 계산될 때까지 온-칩 메모리 및/또는 내부 메모리에 유지될 수 있다.

[0390] 적어도 하나의 메모리는 온-칩 메모리 및/또는 내부 메모리를 포함하고, 동작들은 다음 세트의 가중치, 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분을 온-칩 메모리 및/또는 내부 메모리에 저장하는 단계 및 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분과 다음 세트의 가중치를 계산하는 단계를 포함할 수 있다.

[0391] 상기 동작들은 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치로부터 계산된 값들을 적어도 하나의 메모리에 저장하는 단계; 다음 프로세싱 단계를 위해 적어도 하나의 메모리에 다음 세트의 가중치들을 저장하는 단계; 그리고 계산된 값들과 다음 세트의 가중치를 계산하는 단계를 포함할 수 있다.

[0392] 적어도 하나의 메모리는 내부 메모리를 포함하고, 제 1 값들과 다음 세트의 가중치를 계산한 제 2 값들은 내부 메모리에서만 상주할 수 있다.

[0393] 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부는 완전한 데이터 세트를 포함할 수 있다.

[0394] 적어도 하나의 메모리는 내부 메모리를 포함하고, 동작들은, 내부 메모리에 저장하기 전에, 일 세트의 가중치 값들의 크기, 제 1 배치채널의 적어도 일부의 크기 및 제 2 배치채널의 적어도 일부의 크기 상기를 내부 메모리에 피팅(fitting)되도록 타일링(tiling)하는 단계를 포함할 수 있다.

[0395] 상기 동작들에 대해, ANN은 복수의 배치채널들로부터의 객체 검출, 분류 또는 세그먼트화를 포함하는 적어도 하나의 동작을 수행하도록 구성될 수 있다.

[0396] 상기 동작들을 위해, 객체는 차량, 신호등, 장애물, 통행인, 사람, 동물, 도로, 표지판, 및 도로 선 중 적어도 하나를 포함할 수 있다.

[0397] 상기 ANN에 대한 동작들 전에 복수의 배치채널들을 전처리하는 단계를 포함할 수 있다.

[0398] 상기 동작들을 위해, ANN은 객체들의 향상된 검출을 위해 복수의 배치채널들을 전처리하면서, 복수의 배치채널들로부터 객체들을 동시에 검출하도록 구성될 수 있다.

[0399] 상기 동작들에 대해, 복수의 배치채널들 각각은 복수의 이미지들 각각에 대응할 수 있다.

[0400] 상기 동작들을 위해, 복수의 배치채널들 중 적어도 하나의 배치채널은 IR, RGB, YCBCR, HSV, 및 HIS 형식일 수 있다.

[0401] 상기 동작들을 위해, 복수의 배치채널들 중 적어도 하나는 차량의 내부를 캡처하는 이미지를 포함하고, ANN은 차량의 안전 관련 객체, 기능, 운전자의 상태, 및 승객의 상태 중 적어도 하나를 검출하도록 구성될 수 있다.

[0402] 상기 동작들을 위해, 복수의 이미지들은 RGB 이미지, IR 이미지, RADAR 이미지, ULTRASOUND 이미지, LIDAR 이미지, 열 화상 이미지, NIR 이미지 및 이들의 융합 이미지 중 적어도 하나를 포함할 수 있다.

[0403] 상기 동작들을 위해, 복수의 이미지들은 실질적으로 동일한 기간에 캡처된 이미지들이다.

[0404] 상기 동작들을 위해, 복수의 배치채널들 각각은 복수의 센서 데이터 각각에 대응하고, 그리고 복수의 센서 데이터는 압력 센서, 피에조 센서, 습도 센서, 먼지 센서, 스모그 센서, 소나 센서, 진동 센서, 가속도 센서 또는 모션 센서 중 하나 이상으로부터의 데이터를 포함할 수 있다.

[0405] 본 개시에 따른 신경 프로세싱 유닛은 상기 제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크에 대한 신경 프로세싱 유닛으로서, 제 1 배치채널의 적어도 일부, 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치 값들을 저장하도록 구성된 적어도 하나의 내부 메모리; 그리고 저장된 일 세트의 가중치 값들을 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element)를 포함할 수 있다.

[0406] 적어도 하나의 내부 메모리에 할당되는 제 1 배치채널의 적어도 일부의 크기 및 적어도 제 2 배치채널의 크기는 실질적으로 동일할 수 있다.

[0407] 일 세트의 가중치는 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응할 수 있다.

[0408] 복수의 배치채널들은 제 3 배치채널 및 제 4 배치채널을 포함하고, 적어도 하나의 내부 메모리는, 일 세트의 가중치를 유지하는 동안, 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부를 저장하도록 더 구성되고, PE는 제 3 배치채널의 적어도 일부 및 제 4 배치채널의 적어도 일부 각각과 상기 일 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0409] 적어도 하나의 내부 메모리는 복수의 배치채널들이 계산될 때까지 일 세트의 가중치를 유지하도록 더 구성될 수 있다.

[0410] 적어도 하나의 내부 메모리는 또 다른 세트의 가중치, 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분을 저장하도록 더 구성되고; 그리고 PE는 제 1 배치채널의 다음 부분 및 제 2 배치채널의 다음 부분 각각과 또 다른 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0411] 적어도 하나의 내부 메모리는 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부와 일 세트의 가중치로부터 계산된 값들을 저장하고, 그리고 다음 스테이지에 대한 일 세트의 가중치를 저장하도록 더 구성되고, PE는 계산된 값들과 다음 스테이지에 대한 일 세트의 가중치를 계산하도록 더 구성되고, 일 세트의 가중치는 복수의 배치채널들이 계산될 때까지 내부 메모리에 유지하도록 더 구성될 수 있다.

[0412] 적어도 하나의 내부 메모리는, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 대응하고, 제 1 배치채널의 적어도 일부 및 상기 제 2 배치채널의 적어도 일부와 상기 일 세트의 가중치로부터 계산된 제 1 계산 값들을 저장하고, 다음 프로세싱 단계를 위한 다음 세트의 가중치를 저장하도록 더 구성되고, PE는 제 1 계산 값들과 다음 세트의 가중치를 계산하도록 더 구성될 수 있다.

[0413] 신경 프로세싱 유닛은 일 세트의 가중치의 크기, 제 1 배치채널의 적어도 일부의 크기 및 제 2 배치채널의 적어도 일부의 크기를 내부 메모리에 맞게 조정하도록 구성된 스케줄러를 더 포함할 수 있다.

[0414] 본 개시에 따른 신경 프로세싱 유닛은 제 1 배치채널 및 제 2 배치채널을 포함하는 복수의 배치채널들을 프로세싱하기 위한 인공 신경 네트워크 (ANN) 를 위한 신경 프로세싱 유닛으로서, 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부, 및 일 세트의 가중치 값들을 저장하도록 구성된 적어도 하나의 내부 메모리 및 저장된 일 세트의 가중치 값들을 제 1 배치채널의 적어도 일부 및 제 2 배치채널의 적어도 일부에 적용하도록 구성된 적어도 하나의 PE (processing element) 를 포함하고, 제 1 배치채널의 적어도 일부의 크기는, 적어도 하나의 내부 메모리의 크기를 복수의 채널들의 수로 나눈 것과 같거나 작을 수 있다.

[0415] 적어도 하나의 내부 메모리의 크기는 ANN의 가장 큰 특징맵의 크기 및 배치채널들의 수에 대응할 수 있다.

[0416] 적어도 하나의 내부 메모리는 ANN의 압축된 파라미터들을 저장하도록 더 구성될 수 있다.

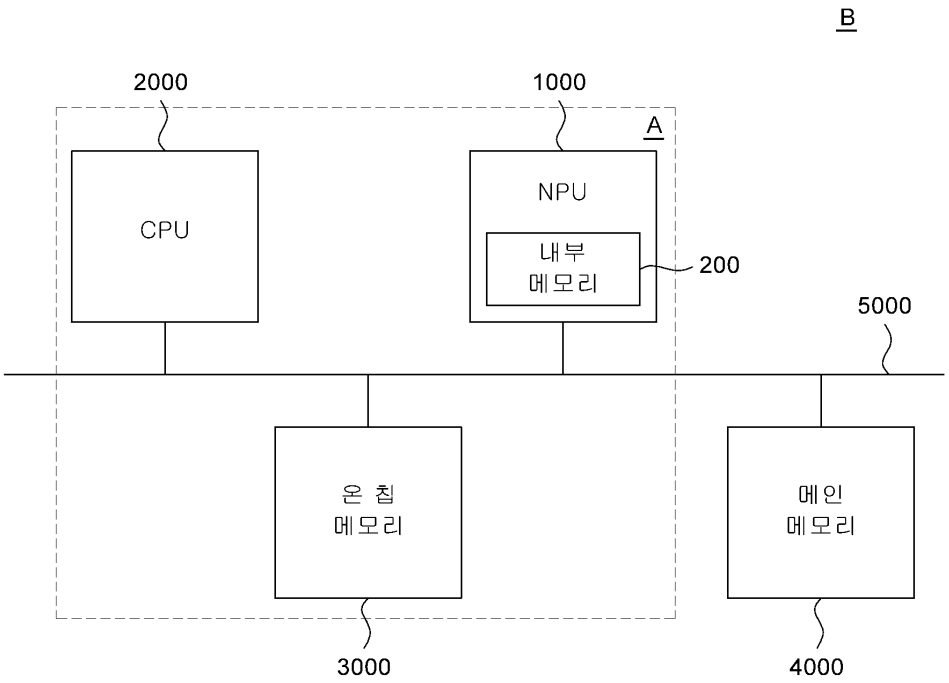
[0417] 적어도 하나의 PE 및 적어도 하나의 내부 메모리와 동작 가능하게 연결되고, 제 1 또는 제 2 배치채널의 적어도 일부의 크기를 조정하도록 구성된 스케줄러를 더 포함할 수 있다.

[0418] 신경 프로세싱 유닛은 적어도 하나의 PE 및 적어도 하나의 내부 메모리 사이에 위치하는 활성화 함수 처리 유닛을 더 포함하고, 활성화 함수 처리 유닛은 제 1 배치채널 및 제 2 배치채널에 대응하는 특징맵들을 순차적으로 처리하여 제 1 배치채널 및 제 2 배치채널에 대응하는 활성화맵들을 순차적으로 출력할 수 있다.

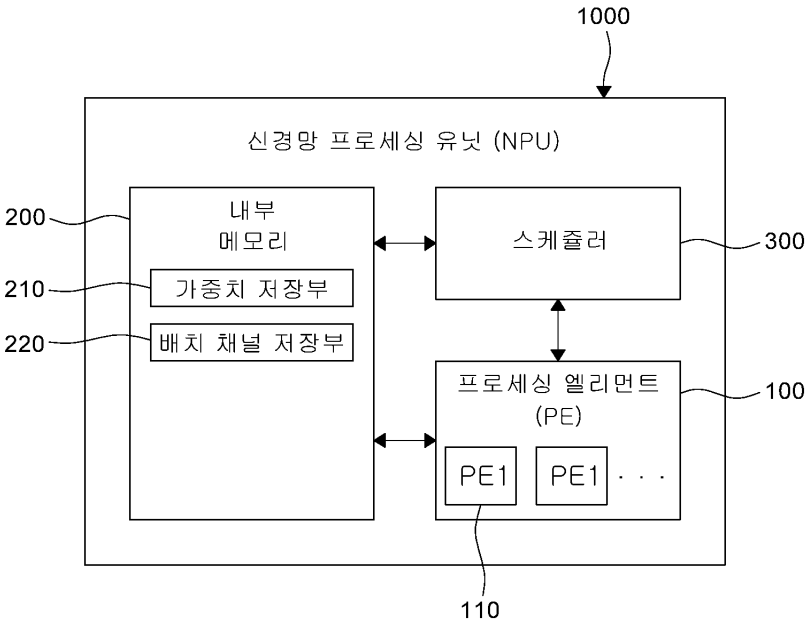
[0419] 본 명세서와 도면에 게시된 본 개시의 예시들은 본 개시의 기술내용을 쉽게 설명하고 본 개시의 이해를 돕기 위해 특정 예를 제시한 것뿐이며, 본 명의 범위를 한정하고자 하는 것은 아니다. 여기에 게시된 예시들 이외에도 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은 본 개시가 속하는 기술 분야에서 통상의 지식을 가진 자에게 자명한 것이다.

도면

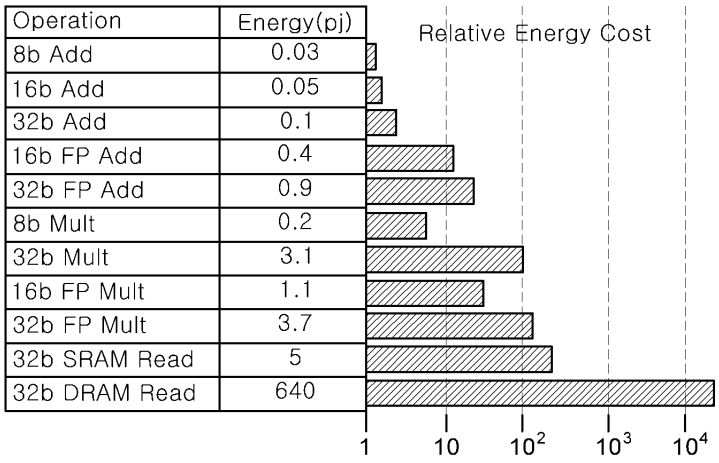
도면1


B
2000
1000
A
CPU
NPU
내부
메모리
200
5000
온 칩
메모리
메인
메모리
3000
4000

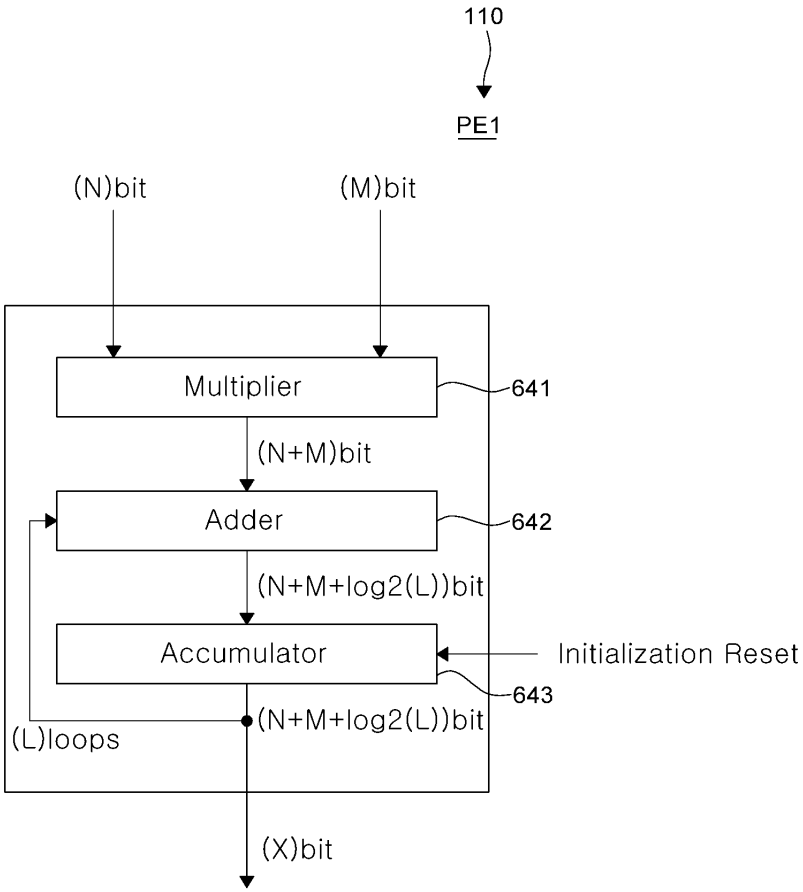
도면2a


1000
신경망 프로세싱 유닛 (NPU)
200
내부
메모리
210
가중치 저장부
220
배치 채널 저장부
스케쥴러
300
프로세싱 엘리먼트
(PE)
100
PE1
PE1
110

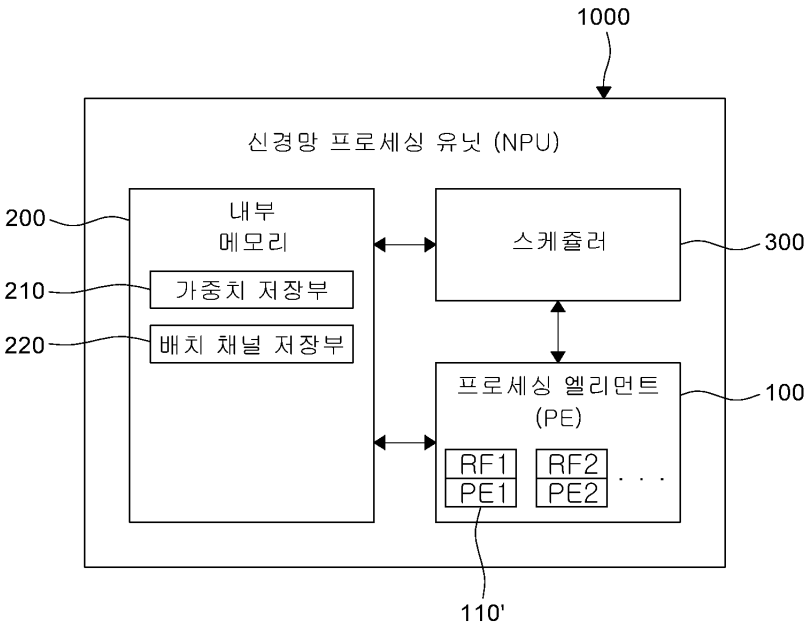
도면2b


Operation
Energy(pj)
8b Add
0.03
16b Add
0.05
32b Add
0.1
16b FP Add
0.4
32b FP Add
0.9
8b Mult
0.2
32b Mult
3.1
16b FP Mult
1.1
32b FP Mult
3.7
32b SRAM Read
5
32b DRAM Read
640
Relative Energy Cost
1
10
10^2
10^3
10^4

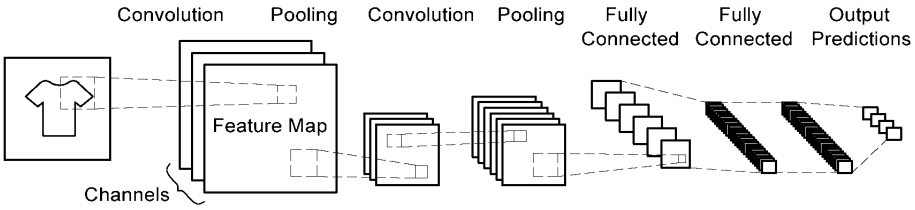
도면2c


110
PE1
(N)bit
(M)bit
Multiplier
641
(N+M)bit
Adder
642
(N+M+log2(L))bit
Accumulator
Initialization Reset
643
(N+M+log2(L))bit
(L)loops
(X)bit

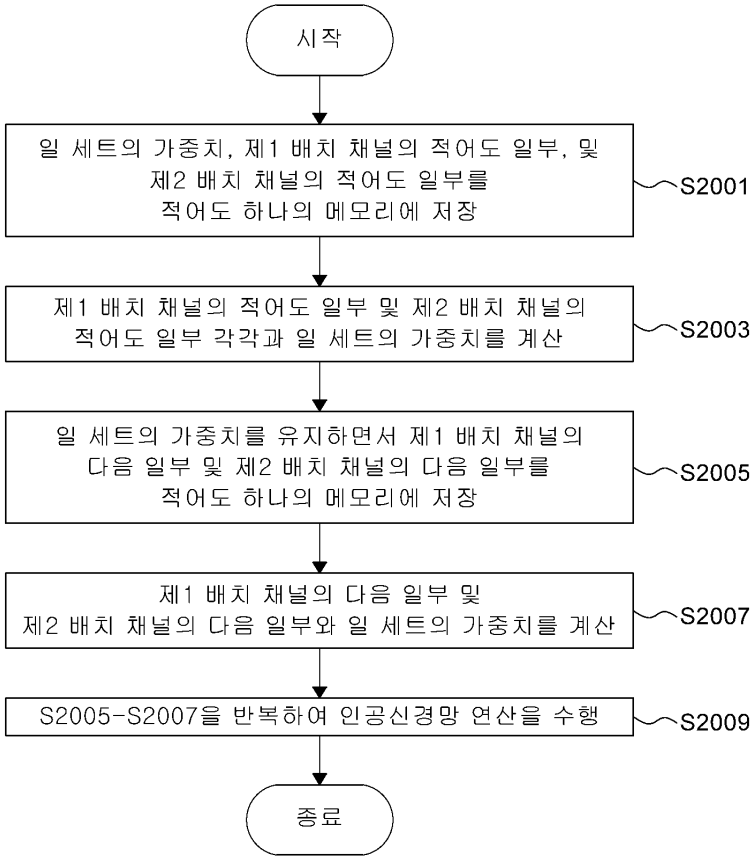
도면3


1000
신경망 프로세싱 유닛 (NPU)
200
내부
메모리
210
가중치 저장부
220
배치 채널 저장부
스케쥴러
300
프로세싱 엘리먼트
(PE)
100
RF1
PE1
RF2
PE2
110'

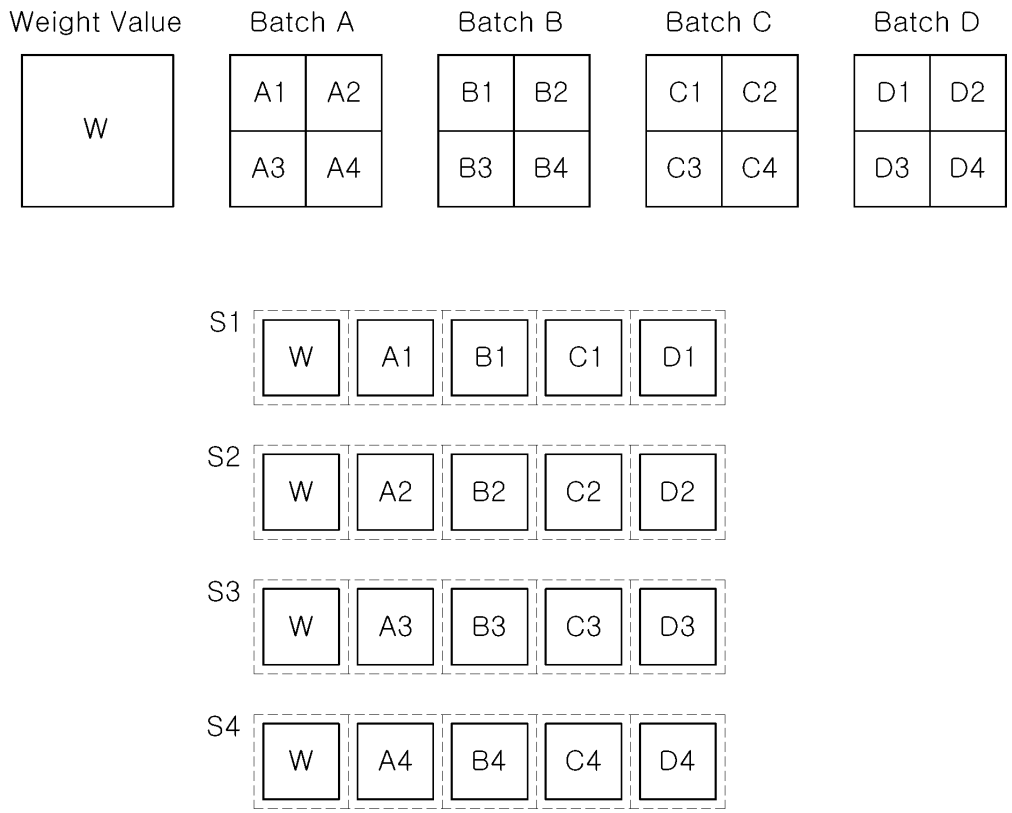
도면4


Convolution
Pooling
Convolution
Pooling
Fully Connected
Fully Connected
Output Predictions
Feature Map
Channels

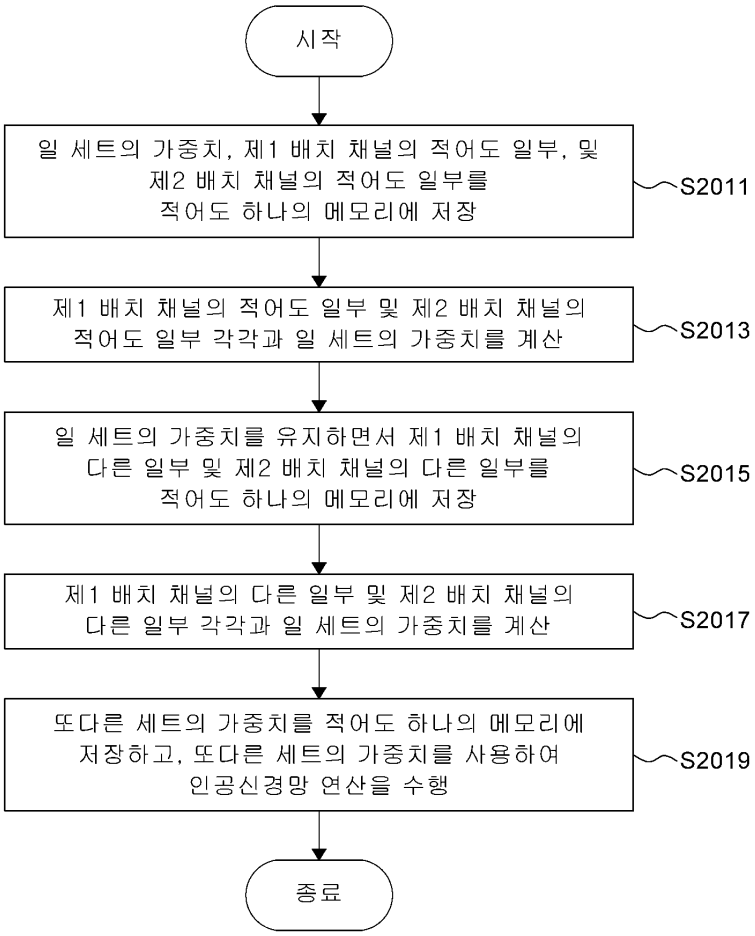
도면5


시작
일 세트의 가중치, 제1 배치 채널의 적어도 일부, 및 제2 배치 채널의 적어도 일부를 적어도 하나의 메모리에 저장
S2001
제1 배치 채널의 적어도 일부 및 제2 배치 채널의 적어도 일부 각각과 일 세트의 가중치를 계산
S2003
일 세트의 가중치를 유지하면서 제1 배치 채널의 다음 일부 및 제2 배치 채널의 다음 일부를 적어도 하나의 메모리에 저장
S2005
제1 배치 채널의 다음 일부 및 제2 배치 채널의 다음 일부와 일 세트의 가중치를 계산
S2007
S2005-S2007을 반복하여 인공신경망 연산을 수행
S2009
종료

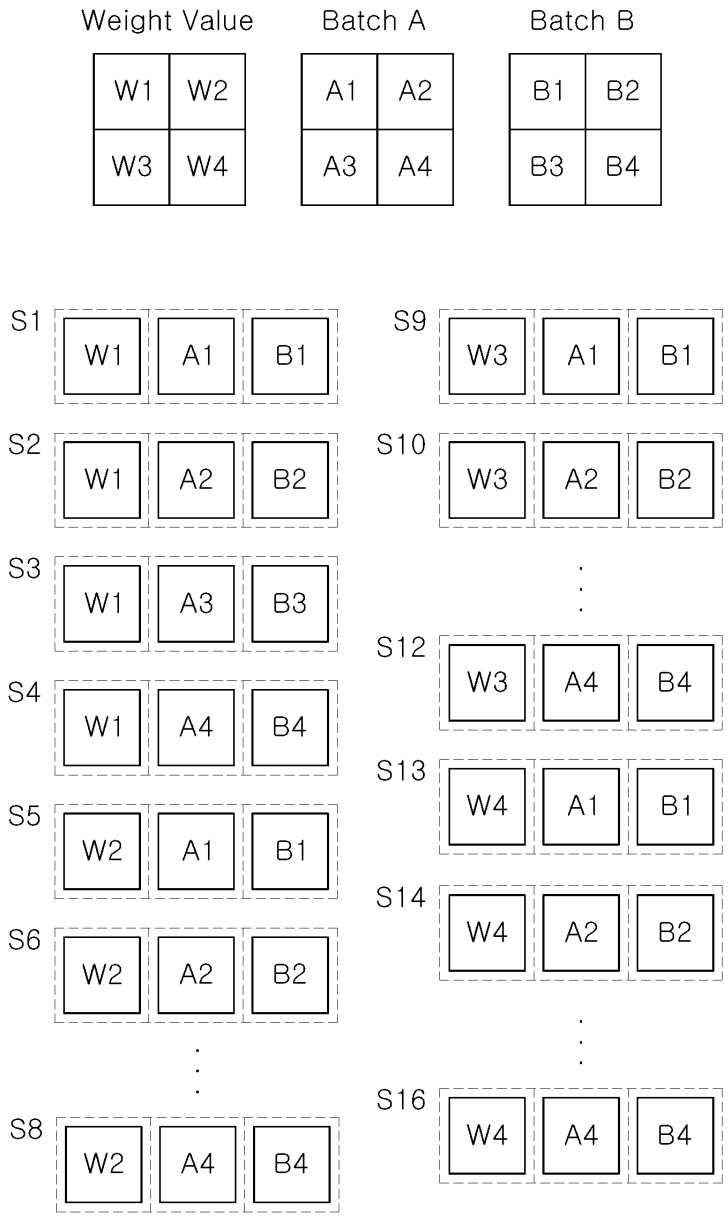
도면6


Weight Value
Batch A
Batch B
Batch C
Batch D
W
A1
A2
A3
A4
B1
B2
B3
B4
C1
C2
C3
C4
D1
D2
D3
D4
S1
W
A1
B1
C1
D1
S2
W
A2
B2
C2
D2
S3
W
A3
B3
C3
D3
S4
W
A4
B4
C4
D4

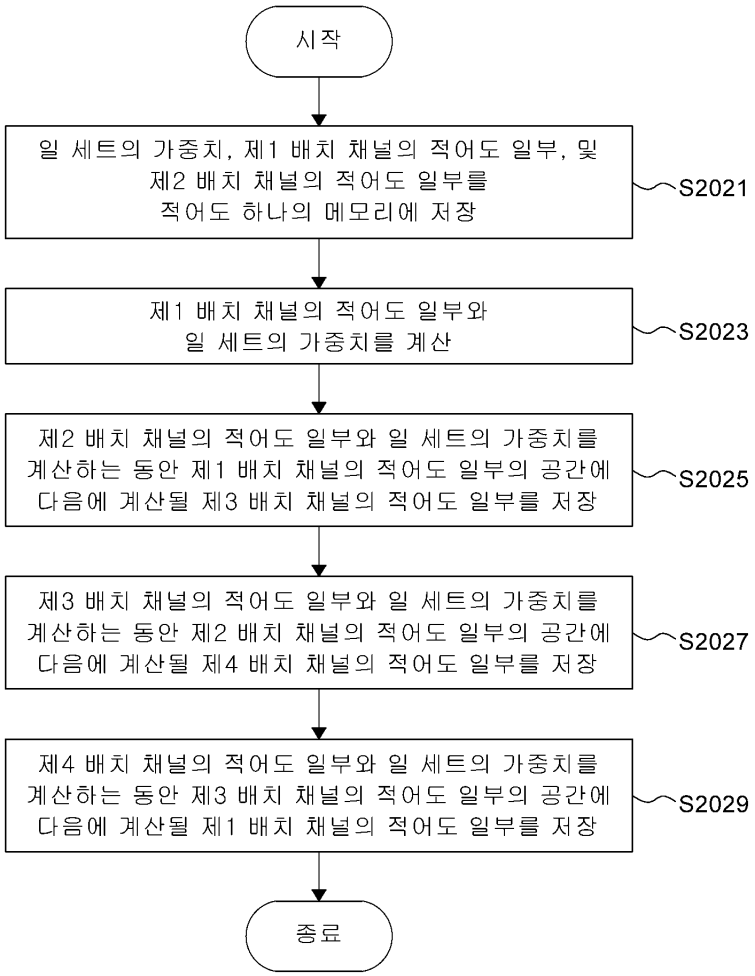
도면7


시작
일 세트의 가중치, 제1 배치 채널의 적어도 일부, 및 제2 배치 채널의 적어도 일부를 적어도 하나의 메모리에 저장
S2011
제1 배치 채널의 적어도 일부 및 제2 배치 채널의 적어도 일부 각각과 일 세트의 가중치를 계산
S2013
일 세트의 가중치를 유지하면서 제1 배치 채널의 다른 일부 및 제2 배치 채널의 다른 일부를 적어도 하나의 메모리에 저장
S2015
제1 배치 채널의 다른 일부 및 제2 배치 채널의 다른 일부 각각과 일 세트의 가중치를 계산
S2017
또다른 세트의 가중치를 적어도 하나의 메모리에 저장하고, 또다른 세트의 가중치를 사용하여 인공신경망 연산을 수행
S2019
종료

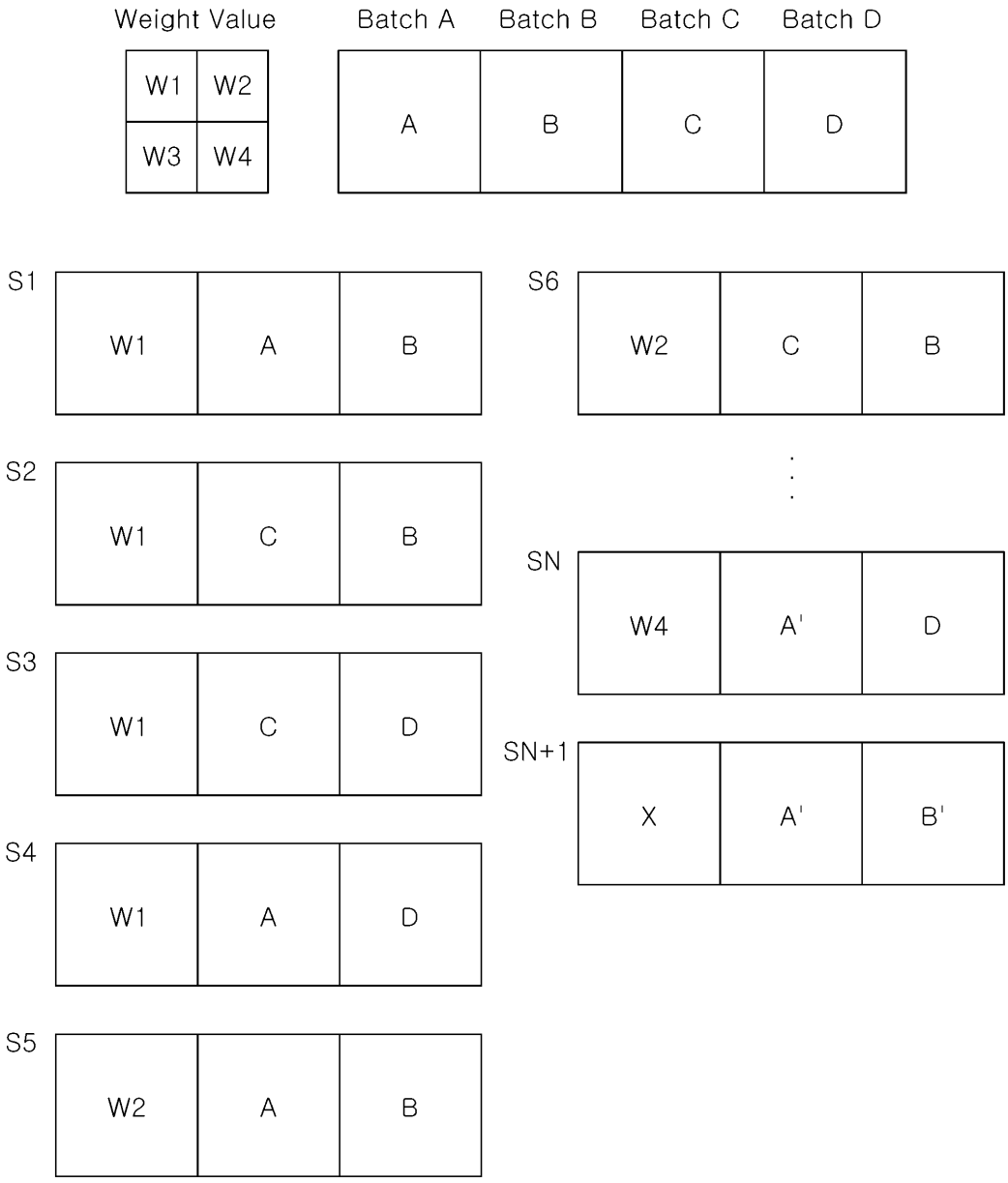
도면8


Weight Value
W1
W2
W3
W4
Batch A
A1
A2
A3
A4
Batch B
B1
B2
B3
B4
S1
W1
A1
B1
S2
W1
A2
B2
S3
W1
A3
B3
S4
W1
A4
B4
S5
W2
A1
B1
S6
W2
A2
B2
S8
W2
A4
B4
S9
W3
A1
B1
S10
W3
A2
B2
S12
W3
A4
B4
S13
W4
A1
B1
S14
W4
A2
B2
S16
W4
A4
B4

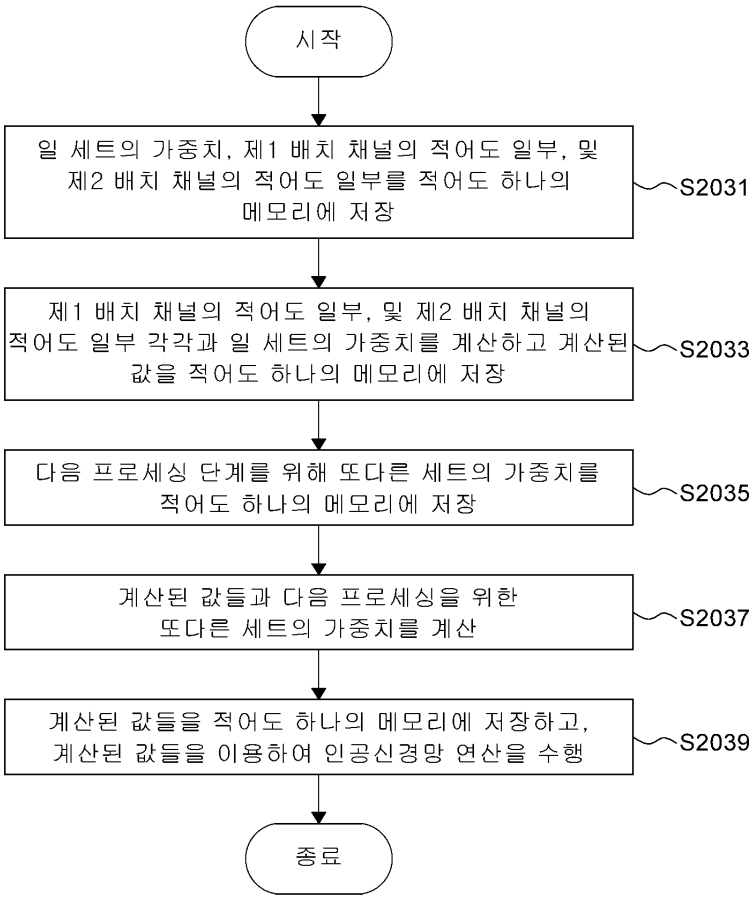
도면9


시작
일 세트의 가중치, 제1 배치 채널의 적어도 일부, 및 제2 배치 채널의 적어도 일부를 적어도 하나의 메모리에 저장
S2021
제1 배치 채널의 적어도 일부와 일 세트의 가중치를 계산
S2023
제2 배치 채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안 제1 배치 채널의 적어도 일부의 공간에 다음에 계산될 제3 배치 채널의 적어도 일부를 저장
S2025
제3 배치 채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안 제2 배치 채널의 적어도 일부의 공간에 다음에 계산될 제4 배치 채널의 적어도 일부를 저장
S2027
제4 배치 채널의 적어도 일부와 일 세트의 가중치를 계산하는 동안 제3 배치 채널의 적어도 일부의 공간에 다음에 계산될 제1 배치 채널의 적어도 일부를 저장
S2029
종료

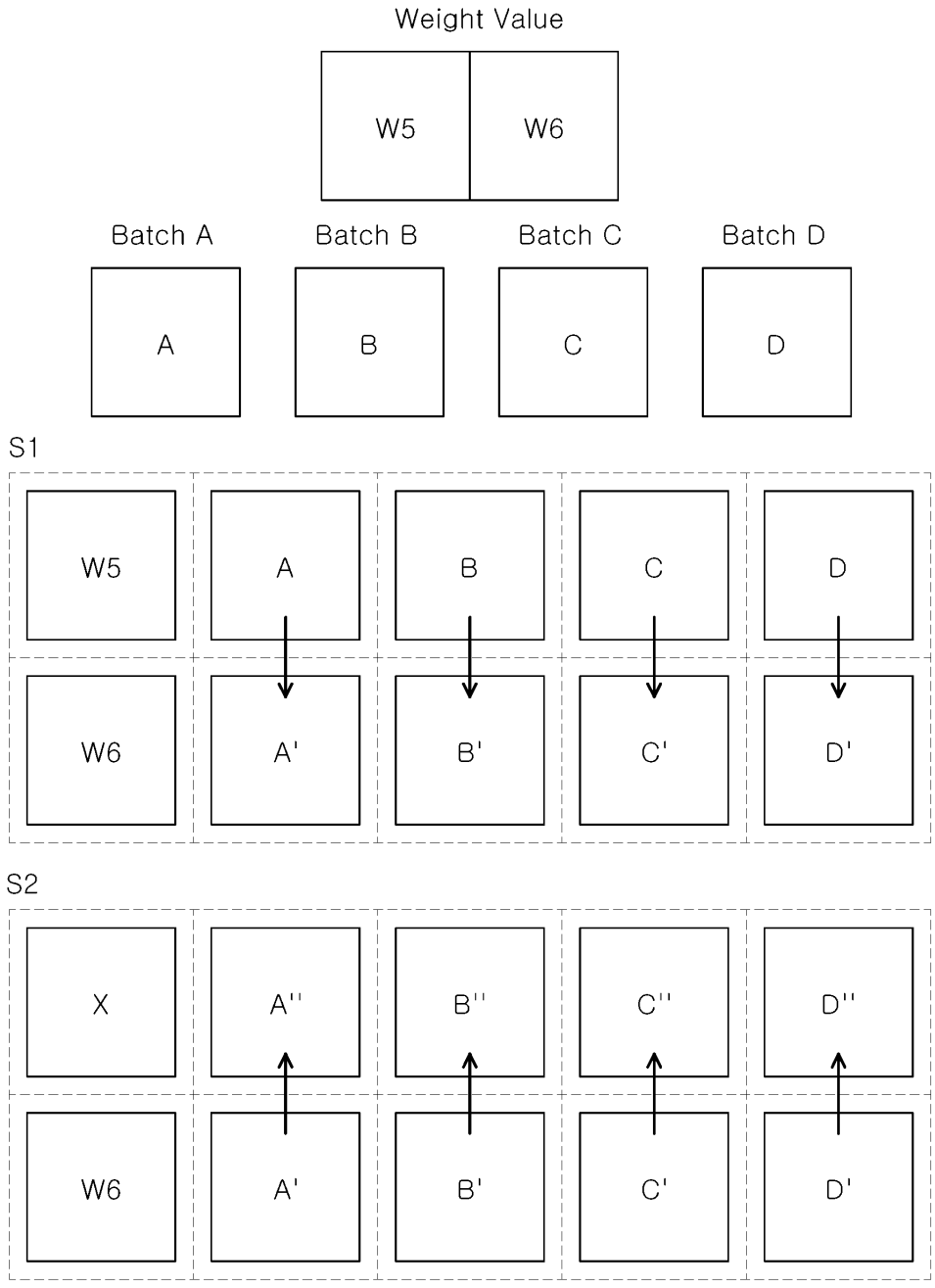
도면10


Weight Value
W1
W2
W3
W4
Batch A
Batch B
Batch C
Batch D
A
B
C
D
S1
W1
A
B
S2
W1
C
B
S3
W1
C
D
S4
W1
A
D
S5
W2
A
B
S6
W2
C
B
SN
W4
A'
D
SN+1
X
A'
B'

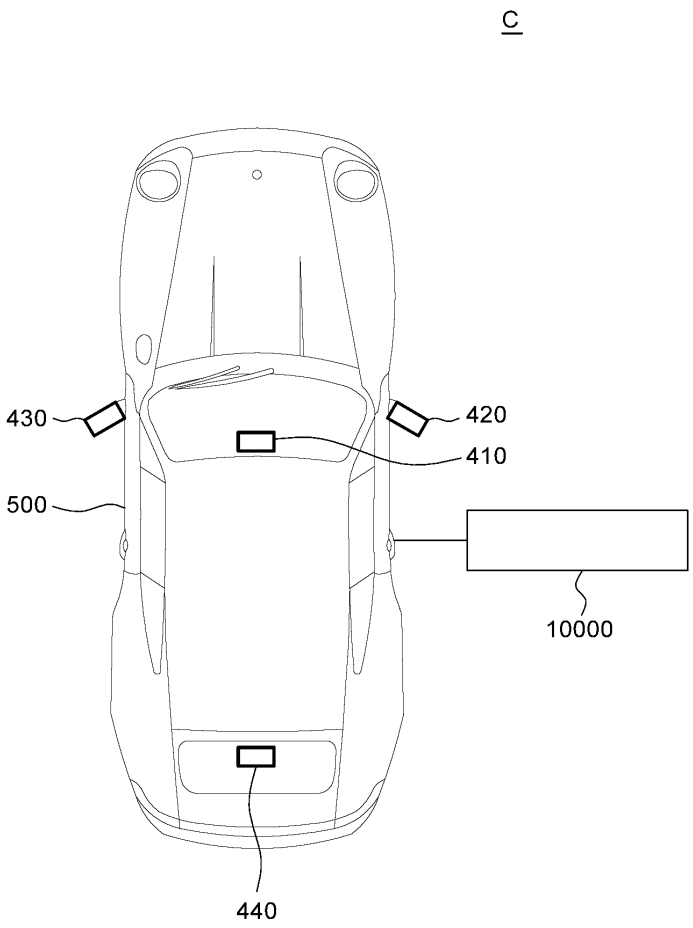
도면11


시작
일 세트의 가중치, 제1 배치 채널의 적어도 일부, 및 제2 배치 채널의 적어도 일부를 적어도 하나의 메모리에 저장
S2031
제1 배치 채널의 적어도 일부, 및 제2 배치 채널의 적어도 일부 각각과 일 세트의 가중치를 계산하고 계산된 값을 적어도 하나의 메모리에 저장
S2033
다음 프로세싱 단계를 위해 또다른 세트의 가중치를 적어도 하나의 메모리에 저장
S2035
계산된 값들과 다음 프로세싱을 위한 또다른 세트의 가중치를 계산
S2037
계산된 값들을 적어도 하나의 메모리에 저장하고, 계산된 값들을 이용하여 인공신경망 연산을 수행
S2039
종료

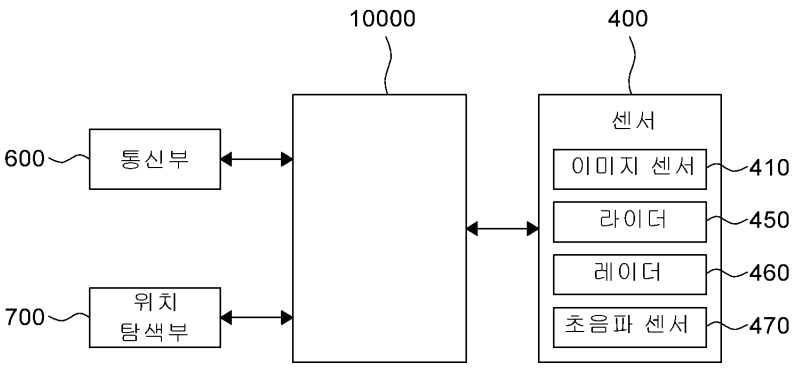
도면12


Weight Value
W5
W6
Batch A
Batch B
Batch C
Batch D
A
B
C
D
S1
W5
A
B
C
D
W6
A'
B'
C'
D'
S2
X
A''
B''
C''
D''
W6
A'
B'
C'
D'

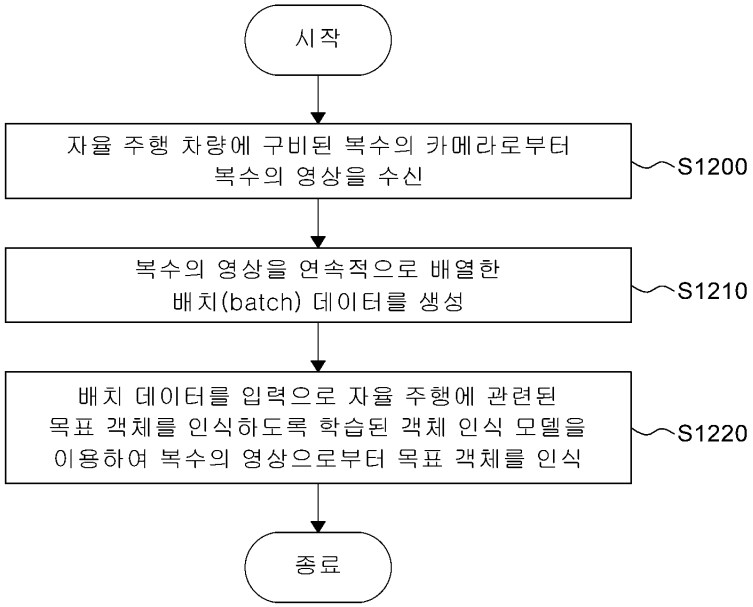
도면13


C
430
420
410
500
10000
440

도면14


10000
400
센서
이미지 센서
410
라이더
450
레이더
460
초음파 센서
470
600
통신부
700
위치
탐색부

도면15


시작
자율 주행 차량에 구비된 복수의 카메라로부터 복수의 영상을 수신
S1200
복수의 영상을 연속적으로 배열한 배치(batch) 데이터를 생성
S1210
배치 데이터를 입력으로 자율 주행에 관련된 목표 객체를 인식하도록 학습된 객체 인식 모델을 이용하여 복수의 영상으로부터 목표 객체를 인식
S1220
종료