(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0025452
(43) 공개일자 2022년03월03일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06F 15/16* (2018.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06F 15/16 (2013.01)
(21) 출원번호 10-2020-0106168
(22) 출원일자 2020년08월24일
심사청구일자 2020년08월24일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
백웅기
울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원
한명균
울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원
(뒷면에 계속)
(74) 대리인
제일특허법인(유)

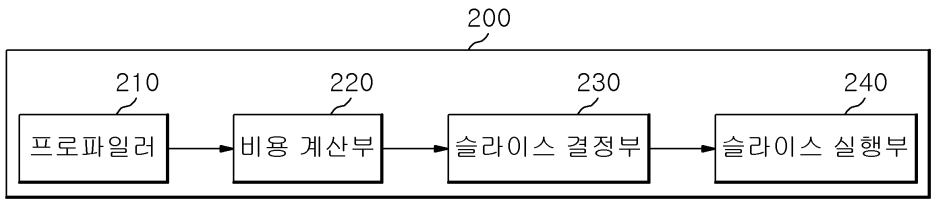
전체 청구항 수 : 총 16 항

(54) 발명의 명칭 **기계학습 모델에 포함된 복수의 레이어들을 분할하고, 분할한 레이어들을 수행하는 프로세서를 결정하는 방법 및 상기 방법을 수행하는 장치**

(57) 요 약

본 발명의 일 실시예에 따른 복수의 프로세서들을 포함하는 장치에서 기계학습 모델을 수행하는 방법은, 복수의 레이어들을 포함하는 기계학습 모델을 입력 받는 단계; 상기 복수의 레이어들 각각의 수행 시간 및 소모 전력 중에서 적어도 하나에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하는 단계; 상기 복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 상기 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정하는 단계; 상기 복수의 레이어들 각각의 상기 프로파일링 데이터, 상기 입력 텐서의 크기 및 상기 출력 텐서의 크기를 이용하여, 상기 복수의 레이어들을 실행할 때의 제1 최소 실행 비용을 계산하는 단계; 상기 제1 최소 실행 비용에 기초하여, 상기 복수의 레이어들을 하나 이상의 슬라이스로 분할하는 단계; 및 상기 복수의 프로세서들 중에서 하나 이상의 프로세서를 이용하여, 상기 하나 이상의 슬라이스를 수행하는 단계를 포함할 수 있다.

대 표 도 - 도2



(72) 발명자

현지훈

울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원

박성범

울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원

박진수

울산광역시 울주군 언양읍 유니스트길 50 울산과학기술원

이 발명을 지원한 국가연구개발사업

과제고유번호	1711107611
과제번호	2018R1C1B6005961
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	대규모 확장 가능한 고성능, 고보안, QoS 보장 빅데이터 처리를 위한 시스템 소프트웨어 기법 연구 및 개발
기 여 율	1/3
과제수행기관명	울산과학기술원
연구기간	2020.03.01 ~ 2021.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103112
과제번호	2018-0-00769-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발(R&D)
연구과제명	인공지능 시스템을 위한 뉴로모픽 컴퓨팅 SW 플랫폼 기술 개발
기 여 율	1/3
과제수행기관명	한국전자통신연구원
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711104673
과제번호	2016M3C4A7952634
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	차세대정보·컴퓨팅기술개발(R&D)
연구과제명	Peta Flops 스케일의 기계학습 프레임워크 개발
기 여 율	1/3
과제수행기관명	울산과학기술원
연구기간	2020.02.01 ~ 2021.06.30

명 세 서

청구범위

청구항 1

복수의 프로세서들을 포함하는 장치에서 기계학습 모델을 수행하는 방법에 있어서,

복수의 레이어들을 포함하는 기계학습 모델을 입력 받는 단계;

상기 복수의 레이어들 각각의 수행 시간 및 소모 전력 중에서 적어도 하나에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하는 단계;

상기 복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 상기 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정하는 단계;

상기 복수의 레이어들 각각의 상기 프로파일링 데이터, 상기 입력 텐서의 크기 및 상기 출력 텐서의 크기를 이용하여, 상기 복수의 레이어들을 실행할 때의 제1 최소 실행 비용을 계산하는 단계;

상기 제1 최소 실행 비용에 기초하여, 상기 복수의 레이어들을 하나 이상의 슬라이스로 분할하는 단계; 및

상기 복수의 프로세서들 중에서 하나 이상의 프로세서를 이용하여, 상기 하나 이상의 슬라이스를 수행하는 단계를 포함하는

기계학습 모델을 수행하는 방법.

청구항 2

제1 항에 있어서,

상기 제1 최소 실행 비용에 기초하여, 상기 복수의 프로세서들 중에서 상기 복수의 레이어들 각각을 수행하는 프로세서를 결정하는 단계를 더 포함하는

기계학습 모델을 수행하는 방법.

청구항 3

제1 항에 있어서,

상기 제1 최소 실행 비용을 계산하는 단계는,

상기 복수의 레이어들 중에서 제1 개수만큼의 연속된 레이어의 프로파일링 데이터, 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제1 개수만큼의 연속된 레이어를 실행할 때의 제1 총 실행 비용, 및 상기 제1 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제2 최소 실행 비용을 결정하는 단계;

상기 복수의 레이어들 중에서 제2 개수만큼의 연속된 레이어의 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제2 개수만큼의 연속된 레이어를 실행할 때의 제2 총 실행 비용, 및 상기 제2 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제3 최소 실행 비용을 결정하는 단계; 및

상기 제1 총 실행 비용 및 상기 제2 최소 실행 비용의 제1 합과 상기 제2 총 실행 비용 및 상기 제3 최소 실행 비용의 제2 합 중에서 더 적은 상기 제1 합을 상기 제1 최소 실행 비용으로 결정하는 단계를 포함하는

기계학습 모델을 수행하는 방법.

청구항 4

제3 항에 있어서,

상기 제1 총 실행 비용은,

상기 제1 개수만큼의 연속된 레이어 중에서 가장 먼저 수행되는 레이어의 입력 텐서의 크기, 상기 제1 개수만큼

의 연속된 레이어 중에서 가장 나중에 수행되는 레이어의 출력 텐서의 크기 및 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용을 이용하여 결정되는

기계학습 모델을 수행하는 방법.

청구항 5

제4 항에 있어서,

상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 프로세서의 성능을 최적화하는 기준에 기초하여 결정되는

기계학습 모델을 수행하는 방법.

청구항 6

제5 항에 있어서,

상기 프로세서의 성능을 최적화하는 기준이 수행 성능인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능에 따라 결정되고,

상기 프로세서의 성능을 최적화하는 기준이 에너지 소모량인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능 및 총 소모 전력에 따라 결정되는

기계학습 모델을 수행하는 방법.

청구항 7

제6 항에 있어서,

상기 수행 성능 및 상기 총 소모 전력은 상기 제1 개수만큼의 연속된 레이어들 각각을 수행하는 프로세서의 동작 주파수에 기초하여 결정되는

기계학습 모델을 수행하는 방법.

청구항 8

복수의 레이어들을 포함하는 기계학습 모델을 수행하는 기계학습 모델 수행 장치에 있어서,

상기 기계학습 모델을 입력 받는 입출력기; 및

복수의 프로세서들을 포함하고, 상기 입출력 장치를 제어하는 프로세서부를 포함하고,

상기 프로세서부는,

상기 복수의 레이어들 각각의 수행 시간 및 소모 전력 중에서 적어도 하나에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하고,

상기 복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 상기 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정하고,

상기 복수의 레이어들 각각의 상기 프로파일링 데이터, 상기 입력 텐서의 크기 및 상기 출력 텐서의 크기를 이용하여, 상기 복수의 레이어들을 실행할 때의 제1 최소 실행 비용을 계산하고,

상기 제1 최소 실행 비용에 기초하여, 상기 복수의 레이어들을 하나 이상의 슬라이스로 분할하고,

상기 복수의 프로세서들 중에서 하나 이상의 프로세서를 이용하여, 상기 하나 이상의 슬라이스를 수행하는

기계학습 모델 수행 장치.

청구항 9

제8 항에 있어서,

상기 프로세서부는,

상기 제1 최소 실행 비용에 기초하여, 상기 복수의 프로세서들 중에서 상기 복수의 레이어들 각각을 수행하는 프로세서를 결정하는

기계학습 모델 수행 장치.

청구항 10

제8 항에 있어서,

상기 프로세서부는,

상기 복수의 레이어들 중에서 제1 개수만큼의 연속된 레이어의 프로파일링 데이터, 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제1 개수만큼의 연속된 레이어를 실행할 때의 제1 총 실행 비용, 및 상기 제1 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제2 최소 실행 비용을 결정하고,

상기 복수의 레이어들 중에서 제2 개수만큼의 연속된 레이어의 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제2 개수만큼의 연속된 레이어를 실행할 때의 제2 총 실행 비용, 및 상기 제2 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제3 최소 실행 비용을 결정하고,

상기 제1 총 실행 비용 및 상기 제2 최소 실행 비용의 제1 합과 상기 제2 총 실행 비용 및 상기 제3 최소 실행 비용의 제2 합 중에서 더 적은 상기 제1 합을 상기 제1 최소 실행 비용으로 결정하는

기계학습 모델 수행 장치.

청구항 11

제10 항에 있어서,

상기 제1 총 실행 비용은,

상기 제1 개수만큼의 연속된 레이어 중에서 가장 먼저 수행되는 레이어의 입력 텐서의 크기, 상기 제1 개수만큼의 연속된 레이어 중에서 가장 나중에 수행되는 레이어의 출력 텐서의 크기 및 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용을 이용하여 결정되는

기계학습 모델 수행 장치.

청구항 12

제11 항에 있어서,

상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 프로세서의 성능을 최적화하는 기준에 기초하여 결정되는

기계학습 모델 수행 장치.

청구항 13

제12 항에 있어서,

상기 프로세서의 성능을 최적화하는 기준이 수행 성능인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능에 따라 결정되고,

상기 프로세서의 성능을 최적화하는 기준이 에너지 소모량인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능 및 총 소모 전력에 따라 결정되는

기계학습 모델 수행 장치.

청구항 14

제13 항에 있어서,

상기 수행 성능 및 상기 총 소모 전력은 상기 제1 개수만큼의 연속된 레이어들 각각을 수행하는 프로세서의 동

작 주파수에 기초하여 결정되는

기계학습 모델 수행 장치.

청구항 15

컴퓨터 프로그램을 저장하고 있는 컴퓨터 판독 가능 기록매체로서,

상기 컴퓨터 프로그램은,

제1 항 내지 제7 항 중 어느 한 항에 따른 방법을 프로세서가 수행하도록 하기 위한 명령어를 포함하는

컴퓨터 판독 가능한 기록매체.

청구항 16

컴퓨터 판독 가능한 기록매체에 저장되어 있는 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은,

제1 항 내지 제7 항 중 어느 한 항에 따른 방법을 프로세서가 수행하도록 하기 위한 명령어를 포함하는

컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 기계학습 모델에 포함된 복수의 레이어들을 분할하고, 분할한 레이어들을 수행하는 프로세서를 결정하는 방법 및 상기 방법을 수행하는 장치에 관한 것이다.

배 경 기 술

[0002] 이종 임베디드 시스템(Heterogeneous embedded system)은 서로 다른 특성들(예컨대, 성능, 에너지, 통신 오버헤드, 기능(실행 가능한 수학 연산), 메모리 용량 등)을 갖는 둘 이상의 프로세서들을 포함하는 시스템이다.

[0003] 이종 임베디드 시스템은 고성능 작업을 수행하며 전력 사용이 큰 빅(big) 코어, 저성능 작업을 수행하며 저전력으로 동작하는 리틀(little) 코어 및 빅 코어와 리틀 코어의 중간 성능의 작업을 수행하는 미들(middle) 코어를 포함하는 CPU(Central Processing Unit), 영상 관련 처리를 중점적으로 수행하는 GPU(Graphics Processing Unit), 머신 러닝 관련 프로세스를 중점적으로 수행하는 NPU(Neural Network Processing Unit)을 포함할 수 있다.

[0004] 최근 기계학습 모델(Machine Learning Model)의 보편화에 따라 임베디드 시스템에서의 기계학습 수행의 필요성이 커지고 있으며, 예컨대, 자율주행 자동차에서의 객체인식, 스마트폰에서 필기체 인식 등 여러 분야에서의 애플리케이션을 실행하기 위해서는 실시간으로 기계학습 추론이 가능해야 한다.

[0005] 다만, 이종 임베디드 시스템에 포함된 프로세서들 중에서 기계학습 모델에 포함되는 각 레이어를 어떠한 프로세서가 수행하느냐에 따라서 기계학습 모델 수행의 성능이 크게 차이가 날 수 있으므로, 기계학습 모델에 포함되는 레이어를 수행할 프로세서를 결정하는 방법이 문제될 수 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명이 해결하고자 하는 과제는, 전술한 문제를 해결하기 위하여, 기계학습 모델에 포함된 복수의 레이어들을 분할하고, 분할한 레이어들을 수행하는 프로세서를 결정하는 방법을 제공하는 것이다.

[0007] 다만, 본 발명이 해결하고자 하는 과제는 이상에서 언급한 것으로 제한되지 않으며, 언급되지 않은 또 다른 해결하고자 하는 과제는 아래의 기재로부터 본 발명이 속하는 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0008] 본 발명의 일 실시예에 따른 복수의 프로세서들을 포함하는 장치에서 기계학습 모델을 수행하는 방법은, 복수의 레이어들을 포함하는 기계학습 모델을 입력 받는 단계; 상기 복수의 레이어들 각각의 수행 시간 및 소모 전력 중에서 적어도 하나에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하는 단계; 상기 복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 상기 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정하는 단계; 상기 복수의 레이어들 각각의 상기 프로파일링 데이터, 상기 입력 텐서의 크기 및 상기 출력 텐서의 크기를 이용하여, 상기 복수의 레이어들을 실행할 때의 제1 최소 실행 비용을 계산하는 단계; 상기 제1 최소 실행 비용에 기초하여, 상기 복수의 레이어들을 하나 이상의 슬라이스로 분할하는 단계; 및 상기 복수의 프로세서들 중에서 하나 이상의 프로세서를 이용하여, 상기 하나 이상의 슬라이스를 수행하는 단계를 포함할 수 있다.

[0009] 상기 방법은, 상기 제1 최소 실행 비용에 기초하여, 상기 복수의 프로세서들 중에서 상기 복수의 레이어들 각각을 수행하는 프로세서를 결정하는 단계를 더 포함할 수 있다.

[0010] 상기 제1 최소 실행 비용을 계산하는 단계는, 상기 복수의 레이어들 중에서 제1 개수만큼의 연속된 레이어의 프로파일링 데이터, 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제1 개수만큼의 연속된 레이어를 실행할 때의 제1 총 실행 비용, 및 상기 제1 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제2 최소 실행 비용을 결정하는 단계; 상기 복수의 레이어들 중에서 제2 개수만큼의 연속된 레이어의 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 상기 복수의 프로세서들 각각이 상기 제2 개수만큼의 연속된 레이어를 실행할 때의 제2 총 실행 비용, 및 상기 제2 개수만큼의 연속된 레이어를 제외한 나머지 레이어에 대한 제3 최소 실행 비용을 결정하는 단계; 및 상기 제1 총 실행 비용 및 상기 제2 최소 실행 비용의 제1 합과 상기 제2 총 실행 비용 및 상기 제3 최소 실행 비용의 제2 합 중에서 더 적은 상기 제1 합을 상기 제1 최소 실행 비용으로 결정하는 단계를 포함할 수 있다.

[0011] 상기 제1 총 실행 비용은, 상기 제1 개수만큼의 연속된 레이어 중에서 가장 먼저 수행되는 레이어의 입력 텐서의 크기, 상기 제1 개수만큼의 연속된 레이어 중에서 가장 나중에 수행되는 레이어의 출력 텐서의 크기 및 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용을 이용하여 결정될 수 있다.

[0012] 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 프로세서의 성능을 최적화하는 기준에 기초하여 결정될 수 있다.

[0013] 상기 프로세서의 성능을 최적화하는 기준이 수행 성능인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능에 따라 결정되고, 상기 프로세서의 성능을 최적화하는 기준이 에너지 소모량인 경우, 상기 제1 개수만큼의 연속된 레이어들 자체의 실행 비용은 상기 1 개수만큼의 연속된 레이어들의 수행 성능 및 총 소모 전력에 따라 결정될 수 있다.

[0014] 상기 수행 성능 및 상기 총 소모 전력은 상기 제1 개수만큼의 연속된 레이어들 각각을 수행하는 프로세서의 동작 주파수에 기초하여 결정될 수 있다.

[0015] 본 발명의 다른 실시예에 따른 복수의 레이어들을 포함하는 기계학습 모델을 수행하는 기계학습 모델 수행 장치는, 상기 기계학습 모델을 입력 받는 입출력기; 및 복수의 프로세서들을 포함하고, 상기 입출력 장치를 제어하는 프로세서부를 포함하고, 상기 프로세서부는, 상기 복수의 레이어들 각각의 수행 시간 및 소모 전력 중에서 적어도 하나에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하고, 상기 복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 상기 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정하고, 상기 복수의 레이어들 각각의 상기 프로파일링 데이터, 상기 입력 텐서의 크기 및 상기 출력 텐서의 크기를 이용하여, 상기 복수의 레이어들을 실행할 때의 제1 최소 실행 비용을 계산하고, 상기 제1 최소 실행 비용에 기초하여, 상기 복수의 레이어들을 하나 이상의 슬라이스로 분할하고, 상기 복수의 프로세서들 중에서 하나 이상의 프로세서를 이용하여, 상기 하나 이상의 슬라이스를 수행할 수 있다.

발명의 효과

[0016] 본 발명의 실시예에 의하면, 기계학습 모델에 포함된 복수의 레이어들을 분할하고, 분할한 레이어들을 수행하는 프로세서를 결정하는 방법을 제안함으로써, 기계학습 모델을 수행하는 소모되는 시간 및 에너지를 효과적으로 감소시킬 수 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 일 실시예에 따라 기계학습 모델의 슬라이스를 각 프로세서로 분배하는 기계학습 모델 수행

장치의 블록도이다.

도 2는 본 발명의 일 실시예에 따른 레이어 분배 모델의 기능을 개념적으로 나타내는 블록도이다.

도 3은 본 발명의 일 실시예에 따른 비용 계산부의 기능을 개념적으로 나타내는 블록도이다.

도 4는 본 발명의 일 실시예에 따라 실행 비용을 결정하는 방법을 나타낸다.

도 5는 본 발명의 일 실시예에 따라 최소 실행 비용을 결정하는 일 단계를 나타낸다.

도 6은 본 발명의 일 실시예에 따라 최소 실행 비용을 결정하는 다른 단계를 나타낸다.

도 7은 도 5 및 도 6의 단계들을 통해 결정된 슬라이스 및 각 슬라이스를 수행하는 프로세서의 일 예시를 나타낸다.

도 8은 본 발명의 일 실시예에 따른 기계학습 모델 수행 장치를 이용하여 기계학습을 실행하는 경우와 다른 종래 기술을 이용하여 기계학습을 실행하는 경우의 레이턴시(latency)를 비교한 예시를 나타낸다.

도 9는 본 발명의 일 실시예에 따른 기계학습 모델 수행 장치를 이용하여 기계학습을 실행하는 경우와 다른 종래 기술을 이용하여 기계학습을 실행하는 경우의 에너지 효율을 비교한 예시를 나타낸다.

도 10은 본 발명의 일 실시예에 따라 복수의 레이어들을 하나 이상의 슬라이스로 분할하고, 각 슬라이스를 수행하는 프로세서를 결정하는 방법을 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0018] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0019] 본 발명의 실시예들을 설명함에 있어서 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다. 그리고 후술되는 용어들은 본 발명의 실시예에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.

[0020] 도 1은 본 발명의 일 실시예에 따라 기계학습 모델의 슬라이스를 각 프로세서로 분배하는 기계학습 모델 수행 장치의 블록도이다.

[0021] 도 1을 참조하면, 기계학습 모델 수행 장치(100)는 프로세서부(110), 입출력기(120) 및 메모리(130)를 포함할 수 있다.

[0022] 프로세서부(110)는 기계학습 모델 수행 장치(100)의 동작을 전반적으로 제어할 수 있다. 프로세서부(110)는 CPU(Central Processing Unit), GPU(Graphics Processing Unit), NPU(Neural Network Processing Unit)를 포함하는 복수의 프로세서들을 포함할 수 있고, CPU는 고성능 작업을 수행하며 전력 사용이 큰 빅(big) 코어, 저성능 작업을 수행하며 저전력으로 동작하는 리틀(little) 코어 및 빅 코어와 리틀 코어의 중간 성능의 작업을 수행하는 미들(middle) 코어를 포함할 수 있다.

[0023] 본 명세서에서 설명하는 프로세서부(110)의 기능은 특별한 설명이 없는 한 프로세서부(110)에 포함된 복수의 프로세서들 중에서 하나 이상의 프로세서가 수행하는 것으로 이해할 수 있다.

[0024] 프로세서부(110)는 입출력기(120)를 제어하여, 기계학습 모델 또는 기계학습 모델을 이용하는 프로그램을 입력받을 수 있다. 또는, 도 1에 도시된 바와는 다르게, 프로세서부(110)는 입출력기(120)가 아닌 송수신기(120)를 통해 기계학습 모델 또는 기계학습 모델을 이용하는 프로그램을 수신할 수 있다. 본 명세서에서는, 설명의 편의를 위하여 기계학습 모델을 입력(수신)받는 경우에 대해서만 설명하지만, 기계학습 모델을 이용하는 프로그램을 수신하는 경우 또한 본 명세서에서 설명하는 방법이 동일하게 적용될 수 있다.

[0025] 메모리(130)는 레이어 분배 모델(200) 및 레이어 분배 모델(200)의 실행에 필요한 정보를 저장할 수 있다.

[0026] 프로세서부(110)는 레이어 분배 모델(200)을 실행하기 위하여 메모리(130)로부터 레이어 분배 모델(200) 및 레

이어 분배 모델(200)의 실행에 필요한 정보를 로드할 수 있다.

[0027] 본 명세서에서 레이어 분배 모델(200)은 입력된 기계학습 모델에 포함된 복수의 레이어들을 하나 이상의 슬라이스로 나누고, 각 슬라이스를 담당할 프로세서를 결정하는 소프트웨어(컴퓨터 프로그램 코드)를 의미할 수 있다.

[0028] 프로세서부(110)는, 레이어 분배 모델(200)을 실행하여, 입력된 기계학습 모델에 포함된 복수의 레이어들을 하나 이상의 슬라이스로 나누고, 복수의 프로세서들 중에서 각 슬라이스를 담당할 프로세서를 결정할 수 있다.

[0029] 도 2는 본 발명의 일 실시예에 따른 레이어 분배 모델의 기능을 개념적으로 나타내는 블록도이고, 도 3은 본 발명의 일 실시예에 따른 비용 계산부의 기능을 개념적으로 나타내는 블록도이고, 도 4는 본 발명의 일 실시예에 따라 실행 비용을 결정하는 방법을 나타낸다.

[0030] 도 2를 참조하면, 레이어 분배 모델(200)은 프로파일러(210), 비용 계산부(220), 슬라이스 결정부(230) 및 슬라이스 실행부(240)를 포함할 수 있다.

[0031] 도 2에 도시된 프로파일러(210), 비용 계산부(220), 슬라이스 결정부(230) 및 슬라이스 실행부(240)는 레이어 분배 모델(200)의 기능을 쉽게 설명하기 위하여 레이어 분배 모델(200)의 기능을 개념적으로 나눈 것으로서, 이에 한정되지 않는다. 실시예들에 따라, 프로파일러(210), 비용 계산부(220), 슬라이스 결정부(230) 및/또는 슬라이스 실행부(240)는 하나의 프로그램에 포함된 일련의 명령어들로 구현될 수도 있다.

[0032] 프로파일러(210)는 입력받은 기계학습 모델에 포함된 복수의 레이어들 각각을 복수의 프로세서들 각각이 수행하는 경우의 수행 시간 및/또는 소모 전력에 대한 프로파일링을 수행하고, 프로파일링 데이터를 생성할 수 있다.

[0033] 상기 프로파일링 데이터는 비용 계산부(220)에서 복수의 프로세서 각각이 복수의 레이어 각각을 수행할 때의 수행 성능을 결정하고, 복수의 프로세서 각각이 복수의 레이어 각각을 수행할 때의 총 소모 전력을 예측하는데 이용될 수 있다.

[0034] 이때, 복수의 프로세서들 중에서 DVFS (Dynamic Voltage and Frequency Scaling)를 지원하는 프로세서가 있는 경우, 프로파일러(210)는 해당 프로세서가 최소 주파수로 레이어를 수행할 때의 수행 시간 및 소모 전력과 해당 프로세서가 최대 주파수로 레이어를 수행할 때의 수행 시간 및 소모 전력을 프로파일링 할 수 있다.

[0035] 또한, 복수의 프로세서들 중에서 특정 프로세서가 메모리의 제한, 실행 가능한 연산의 제한 등으로 인하여 특정 레이어를 수행할 수 없는 경우, 프로파일러(210)는 상기 특정 프로세서가 상기 특정 레이어를 수행할 때의 실행 비용을 무한대로 설정할 수 있다.

[0036] 또한, 프로파일러(210)는, 통신 비용을 결정하는데 이용하기 위하여, 복수의 레이어들 각각에 입력되는 입력 텐서(tensor)의 크기 및 복수의 레이어들 각각으로부터 출력되는 출력 텐서(tensor)의 크기를 결정할 수 있다.

[0037] 프로파일러(210)는 입력 텐서, 출력 텐서의 크기 및 프로파일링 데이터를 비용 계산부(220)로 전송할 수 있다.

[0038] 비용 계산부(220)는 프로파일러(210)로부터 복수의 레이어들 각각의 입력 텐서(tensor)의 크기 및 출력 텐서(tensor)의 크기를 수신할 수 있다.

[0039] 비용 계산부(220)는, 수신한 복수의 레이어들 각각의 입력 텐서(tensor)의 크기 및 출력 텐서(tensor)의 크기를 이용하여, 복수의 프로세서들 각각이 복수의 레이어들 각각을 수행할 때의 통신 비용을 계산할 수 있고, 비용 계산부(220)는, 프로파일러(210)로부터 수신한 프로파일링 데이터를 이용하여, 수행 성능 및 총 소모 전력을 계산할 수 있다.

[0040] 비용 계산부(220)는 계산한 통신 비용, 수행 성능 및 총 소모 전력을 슬라이스 결정부(230)로 전송할 수 있다.

[0041] 도 3을 더 참조하면, 비용 계산부(220)는 비용 추정부(310), 성능 예측부(320) 및 전력 예측부(330)를 포함할 수 있다.

[0042] 도 3에 도시된 비용 추정부(310), 성능 예측부(320) 및 전력 예측부(330)는 비용 계산부(220)의 기능을 쉽게 설명하기 위하여 비용 계산부(220)의 기능을 개념적으로 나눈 것으로서, 이에 한정되지 않는다. 실시예들에 따라, 비용 추정부(310), 성능 예측부(320) 및 전력 예측부(330)는 하나의 프로그램에 포함된 일련의 명령어들로 구현될 수도 있다.

[0043] 비용 추정부(310)는 복수의 레이어들 각각에 대한 입력 텐서(tensor)의 크기(예컨대, Byte 단위)를 이용하여 복수의 프로세서들 각각이 복수의 레이어들 각각을 수행할 때의 입력 통신 비용을 결정하고, 복수의 레이어들 각각에 대한 출력 텐서(tensor)의 크기(예컨대, Byte 단위)를 이용하여 복수의 프로세서들 각각이 복수의 레이어

들 각각을 수행할 때의 출력 통신 비용을 결정할 수 있다.

[0044] 여기서, 입력 통신 비용은 프로세서가 레이어를 수행하기 위해 입력 데이터를 수신하는데 사용하는 자원(또는 자원의 양)(예컨대, 시간, 에너지 등)을 의미하고, 출력 통신 비용은 프로세서가 레이어를 수행함으로 인한 출력 데이터를 전송하는데 사용하는 자원(또는 자원의 양)을 의미할 수 있다.

[0045] 실시예에 따라, 비용 추정부(310)는 선형 회귀(linear regression) 모델에 기초하여 상기 입력 통신 비용 및 상기 출력 통신 비용을 결정할 수 있다. 상기 입력 통신 비용과 상기 출력 통신 비용은 아래의 수학식 1에 기초하여 결정될 수 있다.

수학식 1

[0046]
$$c_{in,l,d,f_d} = \alpha_{d,f_d} \cdot T_{in,l} + \beta_{d,f_d}$$
$$c_{out,l,d,f_d} = \alpha_{d,f_d} \cdot T_{out,l} + \beta_{d,f_d}$$

[0047] 여기서, c_{in}은 입력 통신 비용을 나타내고, l은 레이어(레이어의 인덱스)를 나타내고, d는 프로세서(프로세서의 인덱스)를 나타내고, fd는 d-번째 프로세서의 동작 주파수를 나타내고, c_{out}은 출력 통신 비용을 나타내고, $\alpha_{d,fd}$와 $\beta_{d,fd}$는 선형 회귀 모델에 이용되고 프로세서(d) 및 프로세서(d)의 동작 주파수(fd)마다 선형 회귀를 통해 결정되는 선형 회귀 계수를 나타내고, $T_{in,l}$은 l-번째 레이어에 입력되는 입력 텐서의 크기를 나타내고, $T_{out,l}$은 l-번째 레이어에서 출력되는 출력 텐서의 크기를 나타낼 수 있다.

[0048] 따라서, $c_{in,l,d,fd}$는 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 입력 통신 비용을 의미하고, $c_{out,l,d,fd}$는 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 출력 통신 비용을 의미할 수 있다.

[0049] 성능 예측부(320)는 복수의 프로세서 각각이 복수의 레이어 각각을 수행할 때의 수행 성능을 결정할 수 있다.

[0050] 실시예에 따라, 성능 예측부(320)는 선형 회귀 모델(linear regression model)에 기초하여 수행 성능을 결정할 수 있다. 수행 성능은 아래의 수학식 2에 기초하여 결정될 수 있다.

수학식 2

[0051]
$$t_{l,d,f_d} = \frac{\gamma_{l,d}}{f_d} + \epsilon_{l,d}$$

[0052] 여기서, $\gamma_{l,d}$와 $\epsilon_{l,d}$는 선형 회귀 모델에 이용되고 프로세서(d) 및 레이어(l)마다 프로파일링 데이터 및 선형 회귀를 통해 결정되는 선형 회귀 계수를 나타내고, $t_{l,d,fd}$는 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 수행 성능을 의미할 수 있다.

[0053] 전력 예측부(330)는 복수의 프로세서 각각이 복수의 레이어 각각을 수행할 때의 총 소모 전력을 예측할 수 있다. 상기 총 소모 전력은 프로세서의 동작에 의해 소모되는 동적 전력(dynamic power)과 프로세서의 동작 없이도 기본적으로 소모되는 정적 전력(static power)을 포함할 수 있다.

[0054] 상기 동적 전력은 프로세서의 사양(specification)에 의해 결정되거나, 측정에 의해 결정될 수 있고, 상기 정적 전력은 프로세서 및 프로세서의 동작 주파수에 기초하여 오프라인(offline) 프로파일링을 통해 결정될 수 있다.

[0055] 전력 예측부(330)는 아래의 수학식 3에 기초하여 총 소모 전력을 결정할 수 있다.

수학식 3

$$P_{l,d,f_d} = P_{\text{dynamic},l,d,f_d} + P_{\text{static},d,f_d}$$

$$P_{\text{dynamic},l,d,f_d} = \frac{V_{f_d}^2 \cdot f_d}{V_{f_{d,\max}}^2 \cdot f_{d,\max}} \cdot P_{\text{dynamic},l,d,f_{d,\max}}$$

[0056]

[0057] 여기서, V_{fd}는 동작 주파수가 fd일 때의 프로세서의 동작 전압을 나타내고, $V_{fd,max}$는 DVFS를 지원하는 프로세서의 최대 주파수가 fd,max일 때의 프로세서의 동작 전압을 나타낼 수 있다.

[0058] 따라서, $P_{l,d,fd}$는 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 총 소모 전력을 의미하고, $P_{dynamic,l,d,fd}$는 프로파일러(210)에서 결정한 프로파일링 데이터를 이용하여 결정되는 동적 전력으로서, 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 동적 전력을 의미하고, $P_{static,d,fd}$는 동작 주파수가 fd로 동작하는 d-번째 프로세서가 l-번째 레이어를 수행할 때의 정적 전력을 의미할 수 있다.

[0059] 비용 추정부(310)는 결정한 입력 통신 비용($c_{in,l,d,fd}$), 출력 통신 비용($c_{out,l,d,fd}$), 수행 성능($t_{l,d,fd}$) 및 총 소모 전력($P_{l,d,fd}$)을 슬라이스 결정부(230)로 전송할 수 있다.

[0060] 다시 도 2를 참조하면, 슬라이스 결정부(230)는, 비용 계산부(220)로부터 수신한 입력 통신 비용($c_{in,l,d,fd}$), 출력 통신 비용($c_{out,l,d,fd}$), 수행 성능($t_{l,d,fd}$) 및 총 소모 전력($P_{l,d,fd}$)을 이용하여, 복수의 레이어들을 하나 이상의 슬라이스(slice)로 분할하고, 복수의 프로세서들 중에서 각각의 슬라이스를 수행할 프로세서를 결정할 수 있다. 여기서, 슬라이스란 하나 이상의 연속된 레이어를 포함하는 레이어의 집합을 의미할 수 있다.

[0061] 슬라이스 결정부(230)는 프로세서가 슬라이스(또는 레이어)를 수행하는데 사용하는 자원(또는 자원의 양)(예컨대, 시간, 에너지 등)을 나타내는 실행 비용이 최소가 되도록 복수의 프로세서들에 복수의 레이어들을 슬라이스 단위로 할당할 수 있다. 상기 실행 비용은 실행 시간(단위: s, ms), 에너지(단위: J, mJ) 등으로 나타내어질 수 있다.

[0062] 슬라이스 결정부(230)는, 아래의 수학식 4 및 수학식 5를 이용하여, 복수의 프로세서들 각각에, 복수의 프로세서들 각각이 수행할 슬라이스를 할당할 수 있다.

[0063] 도 4를 더 참조하면, 슬라이스 결정부(230)는, 아래의 수학식 4에 기초하고, 입력 통신 비용, 출력 통신 비용 및 복수의 연속된 레이어들 그 자체를 수행하는 실행 비용을 이용하여, 복수의 연속된 레이어들을 포함하는 슬라이스의 총 실행 비용을 결정할 수 있다.

[0064] 또한, 슬라이스 결정부(230)는, 아래의 수학식 5에 기초하여, 하나 이상의 슬라이스를 수행하는 최소 실행 비용을 결정할 수 있다.

수학식 4

$$C_{m,n,\delta} = \begin{cases} \sum_{k=m}^{n} e_{k,\delta} + c_{in,m,\delta} + c_{out,n,\delta} & \text{If } \delta \text{ can execute} \\ \infty & \text{otherwise} \end{cases}$$

[0065]

수학식 5

$$C_{tot,l} = \begin{cases} 0 & \text{If } l = 0 \\ \min\limits_{0 \le k < l, \forall \delta \in D} (C_{tot,k} + C_{k+1,l,\delta}) & \text{otherwise} \end{cases}$$

[0066]

[0067] 여기서, k는 레이어의 인덱스를 나타내고, $_{\delta}$는 프로세서의 인덱스를 나타낼 수 있다. 따라서, $e_{k,\delta}$는 δ-번째 프로세서가 k-번째 레이어 자체를 수행할 때의 실행 비용을 의미하고, $c_{in,m,\delta}$는 δ-번째 프로세서가 m-번째 레이어를 수행할 때의 입력 통신 비용을 의미하고, $c_{out,n,\delta}$는 δ-번째 프로세서가 n-번째 레이어를 수행할 때의 출력 통신 비용을 의미하고, $C_{m,n,\delta}$는 δ-번째 프로세서가 m-번째부터 n-번째까지의 연속된 레이어들(또는 m-번째부터 n-번째까지의 연속된 레이어들을 포함하는 슬라이스)을 수행할 때의 총 실행 비용을 의미하고, $C_{tot,l}$은 l-개의 연속된 레이어들을 포함하는 슬라이스를 실행하는 최소 실행 비용을 의미할 수 있다.

[0068] 여기서, 최소 실행 비용($C_{tot,l}$)은 복수의 프로세서들 중에서 l-개의 연속된 레이어들을 포함하는 슬라이스를 수행하는 실행 비용이 가장 적은 프로세스가 상기 슬라이스를 수행할 때의 실행 비용을 의미한다.

[0069] 슬라이스 결정부(230)가 슬라이스를 프로세서에 할당하는 기준이 되는 실행 비용은 $c_{tot,l}$를 의미할 수 있다.

[0070] 수학식 5에 나타낸 바와 같이, 최소 실행 비용($C_{tot,l}$)은 k-번째까지의 연속되는 레이어를 포함하는 슬라이스의 최소 실행 비용과 δ-번째 프로세서가 (k+1)-번째부터 l-번째까지의 연속된 레이어들을 수행할 때의 실행 비용의 합이 최소가 될 때의 실행 비용을 의미할 수 있다.

[0071] $e_{k,\delta}$는 수행 성능($t_{l,d,fd}$) 및 총 소모 전력($P_{l,d,fd}$)에 기초하여 결정될 수 있다. $e_{k,\delta}$는 아래의 수학식 6에 기초하여 결정될 수 있다.

수학식 6

$$e_{k,\delta} = \begin{cases} t_{l,d,f_d} & \cdots\cdots\cdots (1) \\ t_{l,d,f_d} \times P_{l,d,f_d} & \cdots\cdots\cdots (2) \end{cases}$$

[0072]

[0073] 수학식 6의 (1)은 기계학습 모델을 수행하는 프로세서부(110)의 성능을 최적화하는 기준이 수행 성능인 경우에 $e_{k,\delta}$를 결정하는 방법을 나타내고, 수학식 6의 (2)는 기계학습 모델을 수행하는 프로세서부(110)의 성능을 최적화 하는 기준이 에너지 소모량인 경우에 $e_{k,\delta}$를 결정하는 방법을 나타낼 수 있다.

[0074] 즉, 기계학습 모델을 수행하는 프로세서부(110)의 성능을 최적화하는 기준이 수행 성능인 경우, δ-번째 프로세서가 k-번째 레이어를 수행할 때의 실행 비용($e_{k,\delta}$)은 수행 성능($t_{l,d,fd}$)으로 결정되고, 기계학습 모델을 수행하는 프로세서부(110)의 성능을 최적화 하는 기준이 에너지 소모량인 경우, δ-번째 프로세서가 k-번째 레이어를 수행할 때의 실행 비용($e_{k,\delta}$)은 수행 성능($t_{l,d,fd}$)과 총 소모 전력($P_{l,d,fd}$)의 곱으로 결정될 수 있다.

[0075] 위에서 살펴본 바와 같이, 슬라이스 결정부(230)는, 입력 통신 비용($c_{in,l,d,fd}$), 출력 통신 비용($c_{out,l,d,fd}$), 수행 성능($t_{l,d,fd}$) 및 총 소모 전력($P_{l,d,fd}$)을 이용하여 결정한 실행 비용에 기초하여, 복수의 레이어들을 하나 이상의 슬라이스로 나누고, 상기 하나 이상의 슬라이스 각각을 복수의 프로세서들 중에서 어느 하나에 할당할 수 있다.

[0076] 보다 상세하게는, 슬라이스 결정부(230)는 슬라이스를 구성하는 연속된 레이어의 개수 및 상기 슬라이스를 수행하는 프로세스를 변경하면서 최소 실행 비용을 계산하고, 계산 결과에 기초하여, 복수의 레이어를 하나 이상의 슬라이스로 분할하고, 복수의 프로세서들 중에서 분할된 슬라이스 각각을 수행할 프로세서를 결정할 수 있다.

슬라이스 결정부(230)가 슬라이스를 분할하고, 분할된 슬라이스를 수행할 프로세서를 결정하는 방법은 도 5 및 도 6을 통해서 보다 자세하게 설명하기로 한다.

[0077] 슬라이스 실행부(240)는 슬라이스 결정부(230)의 결정에 따라 생성된 하나 이상의 슬라이스 각각을 복수의 프로세서들 중에서 각각의 슬라이스를 수행하도록 결정된 프로세서가 실행하도록 제어할 수 있다.

[0078] 도 5는 본 발명의 일 실시예에 따라 최소 실행 비용을 결정하는 일 단계를 나타내고, 도 6은 본 발명의 일 실시예에 따라 최소 실행 비용을 결정하는 다른 단계를 나타내고, 도 7은 도 5 및 도 6의 단계들을 통해 결정된 슬라이스 및 슬라이스를 수행하는 프로세서의 일 예시를 나타낸다.

[0079] 도 5 및 도 6을 참조하면, 기계학습 모델 수행 장치(100)로 입력되는 기계학습 모델은 5개의 레이어들(L1, L2, L3, L4, L5)를 포함할 수 있다. 도 5 및 도 6에서는 설명의 편의를 위하여 기계학습 모델이 5개의 레이어들이 포함되는 것으로 설명하였지만, 이에 한정되지 않는다. 즉, 기계학습 모델에 포함되는 레이어의 수는 변경될 수 있다.

[0080] 슬라이스 결정부(230)는 슬라이스를 구성하는 연속된 레이어의 개수 및 상기 슬라이스를 수행하는 프로세스를 변경하면서 상기 슬라이스에 대한 최소 실행 비용을 계산하고, 계산 결과에 기초하여, 복수의 레이어를 하나 이상의 슬라이스로 분할하고, 복수의 프로세서들 중에서 분할된 슬라이스를 수행할 프로세서를 결정할 수 있다.

[0081] 우선, 도 5를 참조하면, 제1 슬라이스(S1a)가 제5 레이어(L5) 만을 포함한다고 가정하고, 슬라이스 결정부(230)는 복수의 프로세서들 각각이 기계학습 모델에 포함된 5개의 레이어들 중에서 순서적으로 가장 마지막에 수행되는 제5 레이어(L5)를 수행할 때의 총 실행 비용을 계산할 수 있다. 이에 따라, 슬라이스 결정부(230)는 제1 슬라이스(S1a)를 수행하는 실행 비용이 가장 적은 프로세서(δ), 및 실행 비용이 가장 적은 프로세서(δ)가 제1 슬라이스(S1a)를 수행했을 때의 실행 비용을 결정할 수 있다.

[0082] 또한, 슬라이스 결정부(230)는 제5 레이어(L5)를 제외한 나머지 레이어들(L1, L2, L3, L4)을 포함하는 제2 슬라이스(S2a)에 대한 최소 실행 비용을 결정하고, 제1 슬라이스(S1a)에 대한 총 실행 비용($C_{l,1,\delta}$) 및 제2 슬라이스(S2a)에 대한 최소 실행 비용($C_{tot,l-1}$)에 기초하여, 제1 슬라이스(S1a)가 제5 레이어(L5) 하나만을 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다.

[0083] 슬라이스 결정부(230)는, 동일한 방법에 따라, 제1 슬라이스(S1b)가 제4 레이어(L4) 및 제5 레이어(L5)를 포함한다고 가정했을 때, 제1 슬라이스(S1b)를 수행할 프로세서(δ), 제1 슬라이스(S1b)에 대한 총 실행 비용($C_{l-1,1,\delta}$) 및 나머지 레이어들(L1, L2, L3)을 포함하는 제2 슬라이스(S2b)에 대한 최소 실행 비용($C_{tot,l-2}$)을 결정할 수 있다. 슬라이스 결정부(230)는, 결정에 기초하여, 제1 슬라이스(S1b)가 제4 레이어(L4) 및 제5 레이어(L5)를 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다.

[0084] 슬라이스 결정부(230)는, 위의 과정을 반복할 수 있고, 마지막으로 제1 슬라이스(S1c)가 모든 레이어들(L1, L2, L3, L4, L5)을 포함한다고 가정했을 때, 제1 슬라이스(S1c)를 수행할 프로세서(δ) 및 제1 슬라이스(S1c)에 대한 총 실행 비용($C_{l,1,\delta}$)을 결정할 수 있다. 슬라이스 결정부(230)는, 결정에 기초하여, 제1 슬라이스(S1c)가 모든 레이어들(L1, L2, L3, L4, L5)을 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다. 이때, 기계학습 모델 전체의 최소 실행 비용은 제1 슬라이스(S1c)에 대한 총 실행 비용($C_{l,1,\delta}$)과 동일할 수 있다.

[0085] 슬라이스 결정부(230)는 제1 슬라이스(S1a, S1b, S1c, 대표하여 S1)에 포함된 연속된 레이어의 개수에 따라 결정되는 기계학습 모델 전체의 최소 실행 비용을 비교하고, 비교 결과에 따라 제1 슬라이스(S1)에 포함되는 연속된 레이어의 개수 및 제1 슬라이스(S1)를 수행하는 프로세서(δ)를 결정할 수 있다.

[0086] 도 6을 더 참조하면, 도 5에서 설명한 과정에 따라, 제1 슬라이스(S1)가 제5 레이어(L5)만을 포함하는 것으로 결정되었다고 가정한다. 이 경우, 슬라이스 결정부(230)는 제1 슬라이스(S1)로 결정된 제5 레이어(L5)를 제외하고, 다른 레이어들에 대해서 도 5에서 설명한 과정을 반복할 수 있다.

[0087] 즉, 슬라이스 결정부(230)는 제3 슬라이스(S3a)가 제4 레이어(L4)를 포함한다고 가정했을 때, 제3 슬라이스(S3a)를 수행할 프로세서(δ), 제3 슬라이스(S3a)에 대한 총 실행 비용($C_{l-1,l-1,\delta}$) 및 나머지 레이어들(L1, L2, L3)을 포함하는 제4 슬라이스(S4a)에 대한 최소 실행 비용($C_{tot,l-2}$)을 결정할 수 있다. 슬라이스 결정부(230)는, 결정에 기초하여, 제3 슬라이스(S3a)가 제4 레이어(L4)를 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다.

[0088] 슬라이스 결정부(230)는, 동일한 방법에 따라, 제3 슬라이스(S3b)가 제3 레이어(L3) 및 제4 레이어(L4)를 포함한다고 가정했을 때, 제3 슬라이스(S3b)를 수행할 프로세서(δ), 제3 슬라이스(S3b)에 대한 총 실행 비용($C_{l-2,l-1,\delta}$) 및 나머지 레이어들(L1, L2)을 포함하는 제4 슬라이스(S4b)에 대한 최소 실행 비용($C_{tot,l-3}$)을 결정할 수 있다. 슬라이스 결정부(230)는, 결정에 기초하여, 제3 슬라이스(S3b)가 제3 레이어(L3) 및 제4 레이어(L4)를 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다.

[0089] 슬라이스 결정부(230)는, 위의 과정을 반복할 수 있고, 마지막으로 제3 슬라이스(S3c)가 제1 슬라이스(S1)을 제외한 나머지 레이어들(L1, L2, L3, L4)을 포함한다고 가정했을 때, 제3 슬라이스(S3c)를 수행할 프로세서(δ) 및 제3 슬라이스(S3c)에 대한 총 실행 비용($C_{l,l-1,\delta}$)을 결정할 수 있다. 슬라이스 결정부(230)는, 결정에 기초하여, 제3 슬라이스(S3c)가 나머지 레이어들(L1, L2, L3, L4)을 포함했을 때의 기계학습 모델 전체의 최소 실행 비용을 결정할 수 있다. 이때, 기계학습 모델 전체의 최소 실행 비용은 제3 슬라이스(S3c)에 대한 총 실행 비용($C_{l,l-1,\delta}$)과 동일할 수 있다.

[0090] 슬라이스 결정부(230)는 제3 슬라이스(S3a, S3b, S3c, 대표하여 S3)에 포함된 연속된 레이어의 개수에 따라 결정되는 기계학습 모델 전체의 최소 실행 비용을 비교하고, 비교 결과에 따라 제3 슬라이스(S3)에 포함되는 연속된 레이어의 개수 및 제3 슬라이스(S3)를 수행하는 프로세서를 결정할 수 있다.

[0091] 슬라이스 결정부(230)는 도 5 및 도 6을 이용하여 설명한 슬라이스에 포함되는 연속된 레이어의 개수 및 슬라이스를 수행할 프로세서를 결정한 방법을 제1 레이어(L1)가 포함될 슬라이스를 결정할 때까지 반복할 수 있다. 이에 따라, 슬라이스 결정부(230)는 모든 레이어들을 하나 이상의 슬라이스로 분할할 수 있고, 각 슬라이스를 수행할 프로세서를 결정할 수 있다.

[0092] 도 7을 더 참조하면, 기계학습 모델 수행 장치(100)에 이미지를 분석하는 기계학습 모델(LM)이 입력되는 경우, 슬라이스 결정부(230)는 도 5 및 도 6을 이용하여 설명한 방법에 따라, 기계학습 모델(LM)에 포함된 레이어들(L1, L2, L3, L4, L5)을 하나 이상의 슬라이스로 분할하고, 분할한 슬라이스를 수행하는 프로세서를 결정할 수 있다.

[0093] 예컨대, 도 7에 도시된 바와 같이, 슬라이스 결정부(230)는 제1 슬라이스(S1)는 제5 레이어(L1)를 포함하고, CPU의 빅 코어(Big Core)에 의해 수행되도록 결정하고, 제2 슬라이스(S2)는 제3 레이어(L3) 및 제4 레이어(L4)를 포함하고, NPU에 의해 수행되도록 결정하고, 제3 슬라이스(S3)는 제1 레이어(L1) 및 제2 레이어(L2)를 포함하고, GPU에 의해 수행되도록 결정할 수 있다.

[0094] 도 8은 본 발명의 일 실시예에 따른 기계학습 모델 수행 장치를 이용하여 기계학습을 실행하는 경우와 다른 종래 기술을 이용하여 기계학습을 실행하는 경우의 레이턴시(latency)를 비교한 예시를 나타내고, 도 9는 본 발명의 일 실시예에 따른 기계학습 모델 수행 장치를 이용하여 기계학습을 실행하는 경우와 다른 종래 기술을 이용하여 기계학습을 실행하는 경우의 에너지 효율을 비교한 예시를 나타낸다.

[0095] 도 8 및 도 9을 참조하면, MOSAIC은 본 발명의 일 실시예에 따른 기계학습 모델 수행 장치를 이용하여 기계학습 모델을 실행하는 경우를 나타내고, TF-BIG-P, TF-LITTLE-P, TF-GPU-P 및 TF-NPU-P는 각각 최대 주파수로 고정하고 하나의 프로세서(Big Core, Little Core, GPU 또는 NPU) 만을 이용하여 기계학습 모델을 실행하는 경우를 나타내고, TF-BIG-O, TF-LITTLE-O, TF-GPU-O 및 TF-NPU-O는 각각 로드(load)에 따라 동적으로 주파수를 조절되는 하나의 프로세서(Big Core, Little Core, GPU 또는 NPU) 만을 이용하여 기계학습 모델을 실행하는 경우를 나타내고, Exhaustive는 광범위한 실험을 통해 찾은 최적의 설정으로 기계학습 모델을 실행하는 경우를 나타낼 수 있다.

[0096] 도 8을 참조하면, MOSAIC은 TF-BIG-P, TF-LITTLE-P, TF-GPU-P 및 TF-NPU-P와 비교하여 레이턴시(inference latency)가 각각 70.3%, 86.1%, 39.1%, 29.2% 만큼 감소됨을 알 수 있다. 즉, MOSAIC은 heterogeneity, communication, constraint 등을 고려하여 높은 수준의 레이턴시 성능을 달성할 수 있다.

[0097] 또한, MOSAIC은 Exhaustive와 레이턴시가 평균 0.67% 정도 차이가 남을 알 수 있다. 즉, Exhaustive에 따른 최적의 설정을 찾기 위하여 많은 시간 및 자원이 필요한 점을 감안할 때, MOSAIC은 성능 대비 효율이 뛰어남을 알 수 있다.

[0098] 도 9를 참조하면, MOSAIC은 TF-BIG-O, TF-LITTLE-O, TF-GPU-O, TF-NPU-O와 비교하여 사용 에너지(inference energy)가 각각 91.0%, 80.5%, 83.4%, 36.6% 만큼 감소됨을 알 수 있다. 즉, MOSAIC은 heterogeneity,

communication, constraint 등을 고려하여 높은 수준의 에너지 효율을 달성할 수 있다.

[0099] 또한, MOSAIC은 Exhaustive와 레이턴시가 평균 0.53% 정도 차이가 남을 알 수 있다. 즉, Exhaustive에 따른 최적의 설정을 찾기 위하여 많은 시간 및 자원이 필요한 점을 감안할 때, MOSAIC은 성능 대비 효율이 뛰어남을 알 수 있다.

[0100] 도 10은 본 발명의 일 실시예에 따라 복수의 레이어들을 하나 이상의 슬라이스로 분할하고, 각 슬라이스를 수행하는 프로세서를 결정하는 방법을 나타낸다.

[0101] 도 2 내지 도 4 및 도 10을 참조하면, 프로파일러(210)는 기계학습 모델을 입력받고(S1000), 입력받은 기계학습 모델에 포함된 복수의 레이어들 각각을 복수의 프로세서들 각각이 수행하는 경우, 수행 시간 및/또는 소모 전력에 대한 프로파일링을 수행하여 프로파일링 데이터를 생성하고, 복수의 레이어들 각각에 입력되는 입력 텐서(tensor)의 크기 및 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정할 수 있다(S1010).

[0102] 비용 계산부(220)는, 복수의 레이어들 각각의 프로파일링 데이터, 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 복수의 프로세서들 각각이 복수의 레이어들 각각을 수행할 때의 통신 비용, 수행 성능 및 총 소모 전력을 계산할 수 있다(S1020).

[0103] 슬라이스 결정부(230)는, 비용 계산부(220)로부터 수신한 입력 통신 비용, 출력 통신 비용, 수행 성능 및 총 소모 전력을 이용하여, 복수의 레이어들을 하나 이상의 슬라이스로 분할하고, 각 슬라이스를 수행할 프로세서를 결정할 수 있다(S1030).

[0104] 슬라이스 실행부(240)는, 슬라이스 결정부(230)의 결정에 따라 각 슬라이스를 수행하기로 결정된 각 프로세서를 이용하여, 각 슬라이스를 수행할 수 있다(S1040).

[0105] 본 발명에 첨부된 블록도의 각 블록과 흐름도의 각 단계의 조합들은 컴퓨터 프로그램 인스트럭션들에 의해 수행될 수도 있다. 이들 컴퓨터 프로그램 인스트럭션들은 범용 컴퓨터, 특수용 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 인코딩 프로세서에 탑재될 수 있으므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비의 인코딩 프로세서를 통해 수행되는 그 인스트럭션들이 블록도의 각 블록 또는 흐름도의 각 단계에서 설명된 기능들을 수행하는 수단을 생성하게 된다. 이들 컴퓨터 프로그램 인스트럭션들은 특정 방법으로 기능을 구현하기 위해 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 지향할 수 있는 컴퓨터 이용 가능 또는 컴퓨터 판독 가능 메모리에 저장되는 것도 가능하므로, 그 컴퓨터 이용가능 또는 컴퓨터 판독 가능 메모리에 저장된 인스트럭션들은 블록도의 각 블록 또는 흐름도 각 단계에서 설명된 기능을 수행하는 인스트럭션 수단을 내포하는 제조 품목을 생산하는 것도 가능하다. 컴퓨터 프로그램 인스트럭션들은 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에 탑재되는 것도 가능하므로, 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비 상에서 일련의 동작 단계들이 수행되어 컴퓨터로 실행되는 프로세스를 생성해서 컴퓨터 또는 기타 프로그램 가능한 데이터 프로세싱 장비를 수행하는 인스트럭션들은 블록도의 각 블록 및 흐름도의 각 단계에서 설명된 기능들을 실행하기 위한 단계들을 제공하는 것도 가능하다.

[0106] 또한, 각 블록 또는 각 단계는 특정된 논리적 기능(들)을 실행하기 위한 하나 이상의 실행 가능한 인스트럭션들을 포함하는 모듈, 세그먼트 또는 코드의 일부를 나타낼 수 있다. 또, 몇 가지 대체 실시예들에서는 블록들 또는 단계들에서 언급된 기능들이 순서를 벗어나서 발생하는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 사실 실질적으로 동시에 수행되는 것도 가능하고 또는 그 블록들 또는 단계들이 때때로 해당하는 기능에 따라 역순으로 수행되는 것도 가능하다.

[0107] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 품질에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 균등한 범위 내에 있는 모든 기술사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0108] 100: 기계학습 모델 수행 장치

110: 프로세서부

120: 입출력기

130: 메모리

200: 레이어 분배 모델

210: 프로파일러

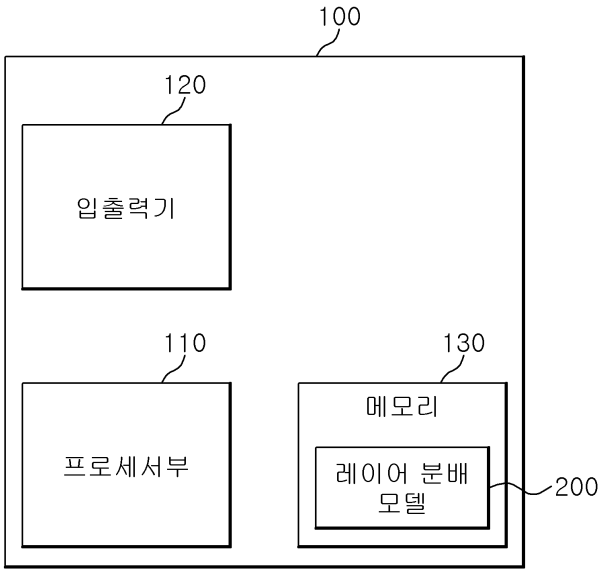
220: 비용 계산부

230: 슬라이스 결정부

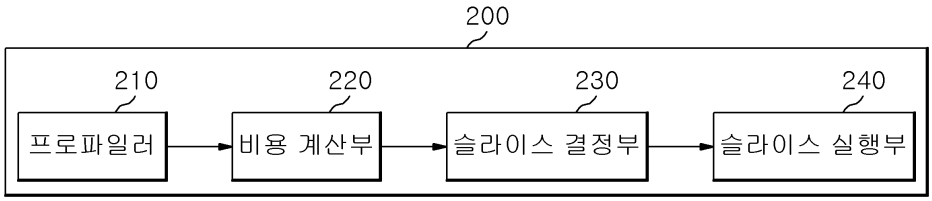
240: 슬라이스 실행부

도면

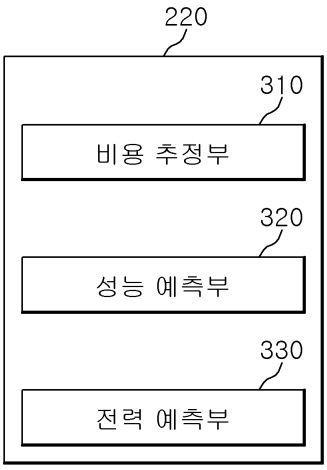
도면1



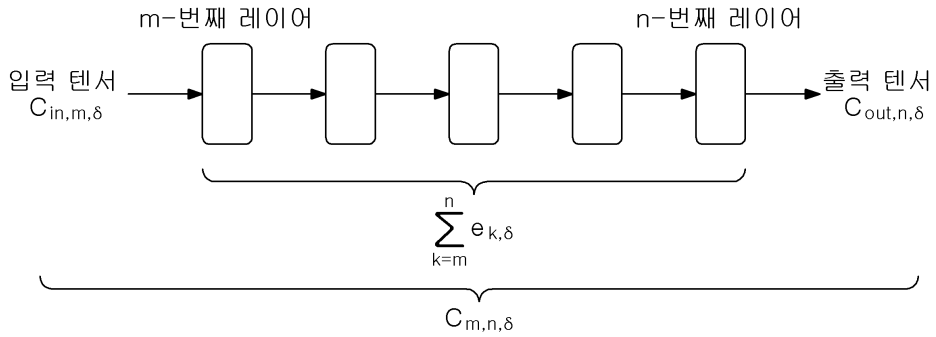
도면2



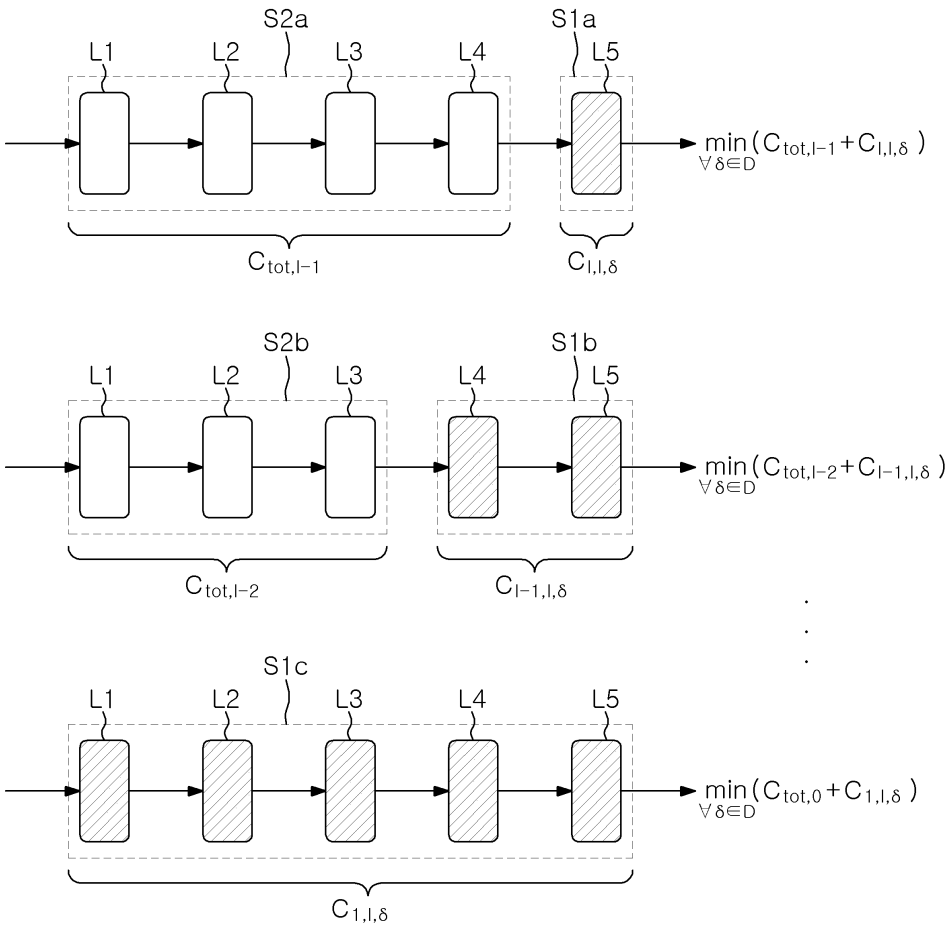
도면3


220
310
비용 추정부
320
성능 예측부
330
전력 예측부

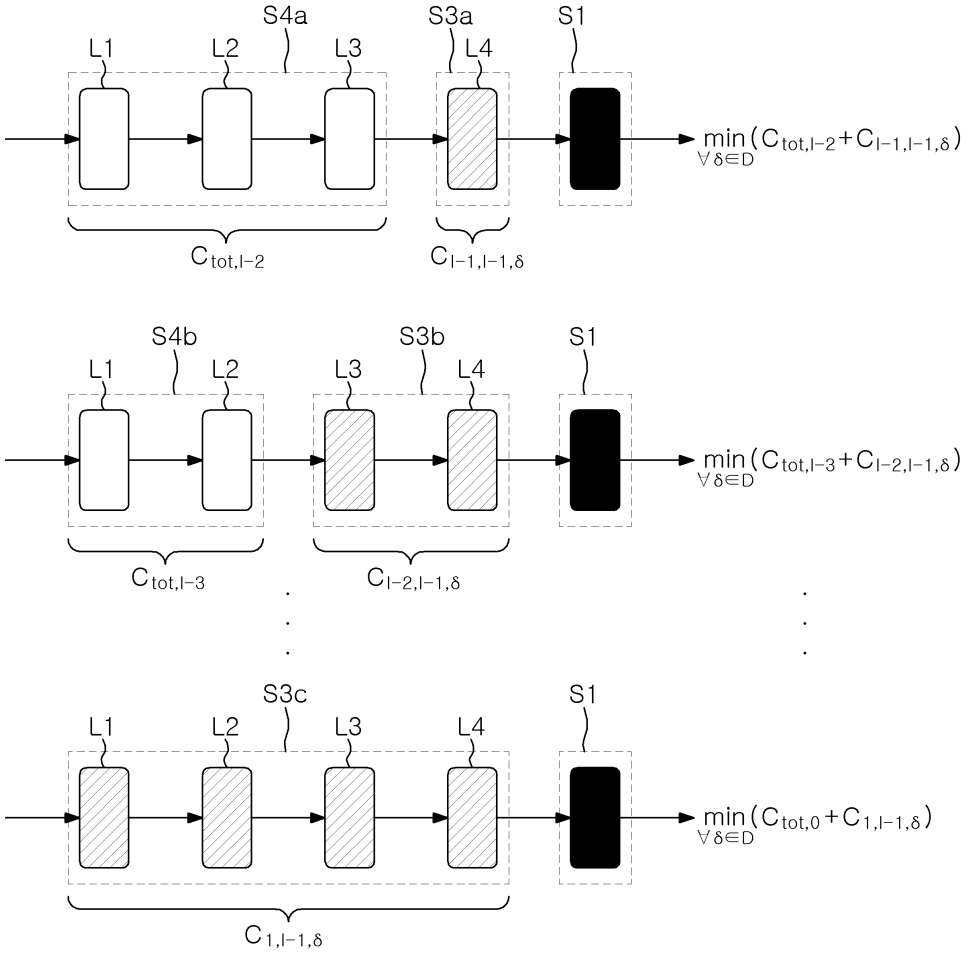
도면4


m-번째 레이어
n-번째 레이어
입력 텐서
$C_{in,m,\delta}$
출력 텐서
$C_{out,n,\delta}$
$\sum_{k=m}^{n} e_{k,\delta}$
$C_{m,n,\delta}$

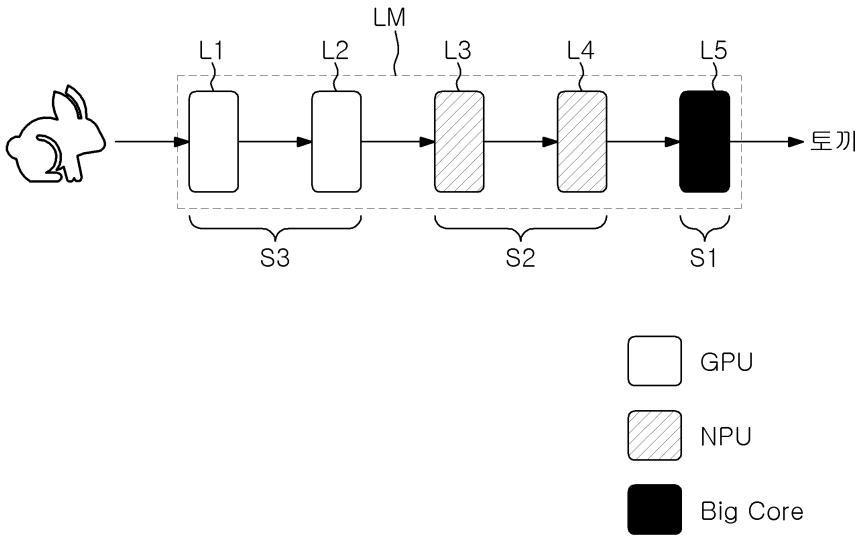
도면5


S2a
S1a
L1
L2
L3
L4
L5
$\min_{\forall\delta\in D}(C_{tot,l-1}+C_{l,l,\delta})$
$C_{tot,l-1}$
$C_{l,l,\delta}$
S2b
S1b
$\min_{\forall\delta\in D}(C_{tot,l-2}+C_{l-1,l,\delta})$
$C_{tot,l-2}$
$C_{l-1,l,\delta}$
S1c
$\min_{\forall\delta\in D}(C_{tot,0}+C_{1,l,\delta})$
$C_{1,l,\delta}$

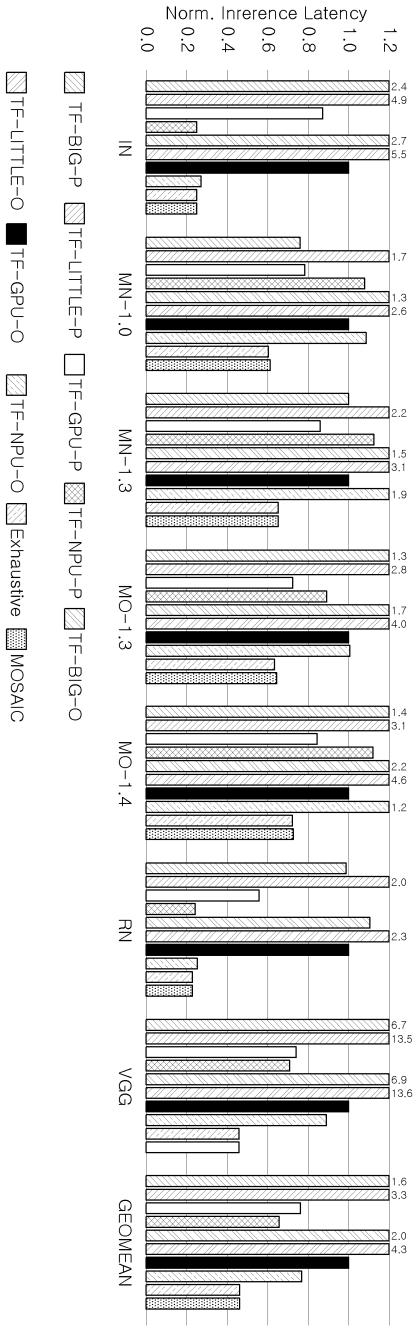
도면6


S4a
S3a
S1
L1
L2
L3
L4
min (C tot,l-2 + C l-1,l-1,δ)
∀δ∈D
C tot,l-2
C l-1,l-1,δ
S4b
S3b
S1
L1
L2
L3
L4
min (C tot,l-3 + C l-2,l-1,δ)
∀δ∈D
C tot,l-3
C l-2,l-1,δ
S3c
S1
L1
L2
L3
L4
min (C tot,0 + C 1,l-1,δ)
∀δ∈D
C 1,l-1,δ

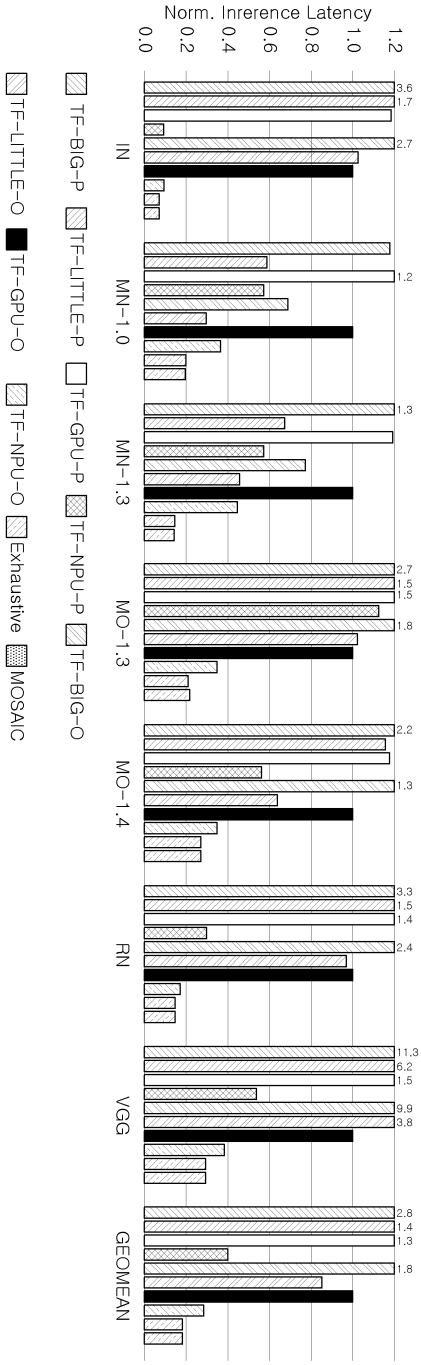
도면7


LM
L1
L2
L3
L4
L5
토끼
S3
S2
S1
GPU
NPU
Big Core

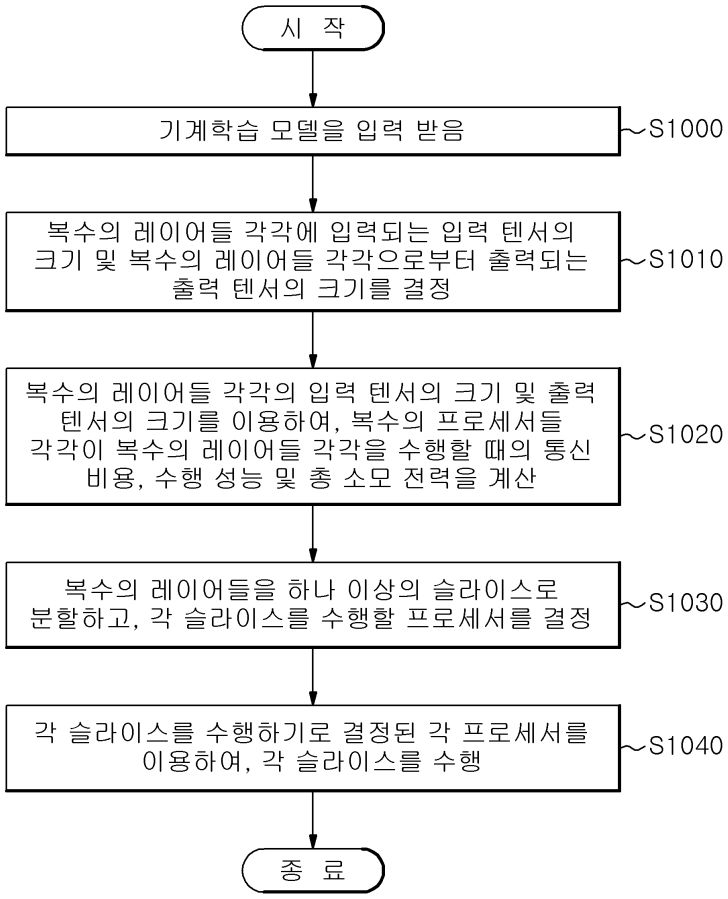
도면8


Norm. Inrerence Latency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
IN
MN-1.0
MN-1.3
MO-1.3
MO-1.4
RN
VGG
GEOMEAN
2.4
4.9
2.7
5.5
1.7
1.3
2.6
2.2
1.5
3.1
1.9
1.3
2.8
1.7
4.0
1.4
3.1
2.2
4.6
1.2
2.0
2.3
6.7
13.5
6.9
13.6
1.6
3.3
2.0
4.3
TF-BIG-P
TF-LITTLE-P
TF-GPU-P
TF-NPU-P
TF-BIG-O
TF-LITTLE-O
TF-GPU-O
TF-NPU-O
Exhaustive
MOSAIC

도면9


Norm. Inrerence Latency
0.0
0.2
0.4
0.6
0.8
1.0
1.2
IN
MN-1.0
MN-1.3
MO-1.3
MO-1.4
RN
VGG
GEOMEAN
3.6
1.7
2.7
1.2
1.3
2.7
1.5
1.5
1.8
2.2
1.3
3.3
1.5
1.4
2.4
11.3
6.2
1.5
9.9
3.8
2.8
1.4
1.3
1.8
TF-BIG-P
TF-LITTLE-P
TF-GPU-P
TF-NPU-P
TF-BIG-O
TF-LITTLE-O
TF-GPU-O
TF-NPU-O
Exhaustive
MOSAIC

도면10


시 작
기계학습 모델을 입력 받음
S1000
복수의 레이어들 각각에 입력되는 입력 텐서의 크기 및 복수의 레이어들 각각으로부터 출력되는 출력 텐서의 크기를 결정
S1010
복수의 레이어들 각각의 입력 텐서의 크기 및 출력 텐서의 크기를 이용하여, 복수의 프로세서들 각각이 복수의 레이어들 각각을 수행할 때의 통신 비용, 수행 성능 및 총 소모 전력을 계산
S1020
복수의 레이어들을 하나 이상의 슬라이스로 분할하고, 각 슬라이스를 수행할 프로세서를 결정
S1030
각 슬라이스를 수행하기로 결정된 각 프로세서를 이용하여, 각 슬라이스를 수행
S1040
종 료