



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0023660
(43) 공개일자 2020년03월06일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/082 (2013.01)
G06N 3/04 (2013.01)
(21) 출원번호 10-2018-0094583
(22) 출원일자 2018년08월13일
심사청구일자 2018년08월13일

(71) 출원인
인천대학교 산학협력단
인천광역시 연수구 아카데미로 119 (송도동)
(72) 발명자
강우철
인천광역시 연수구 아카데미로 119 인천대학교 정보기술대(7호관) 309호
(74) 대리인
김효성

전체 청구항 수 : 총 14 항

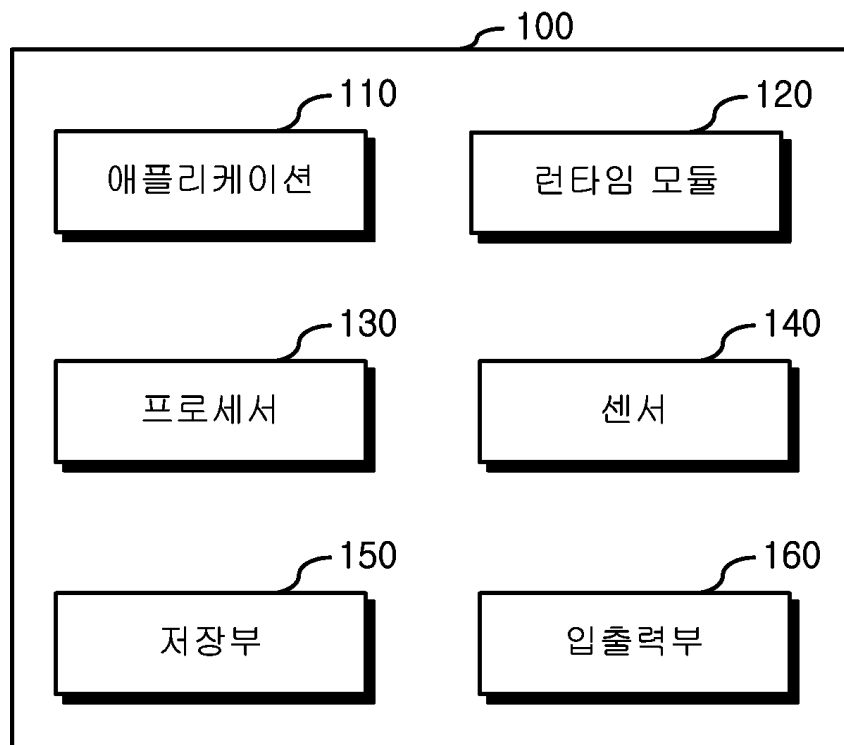
(54) 발명의 명칭 딥러닝 모델을 통한 추론 서비스를 제공할 때, 적어도 하나의 프로세서의 성능을 제어하는 전자 장치 및 그의 동작 방법

(57) 요약

본 발명은 딥러닝 모델을 통한 추론 서비스를 제공할 때, 적어도 하나의 프로세서의 성능을 제어하는 전자 장치 및 그의 동작 방법에 대한 것이다. 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치는, 입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하

(뒷면에 계속)

대표도 - 도1



고, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 레이어 관리부, 상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하고, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 성능 확인부, 상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하고, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 지연 제어부, 상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 리소스 관리부, 및 상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 수행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 수신된 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 입출력부를 포함할 수 있다.

이 발명을 지원한 국가연구개발사업

과제고유번호	NRF-2016R1D1A1B03934266
부처명	교육부
연구관리전문기관	한국연구재단
연구사업명	기본연구지원사업
연구과제명	지능형 모바일/임베디드 시스템을 위한 저전력, 실시간, 경량 딥러닝 SW프레임워크 핵심기술 개발
기 여 율	1/1
주관기관	인천대학교 산학협력단
연구기간	2016.11.01 ~ 2019.10.31

명세서

청구범위

청구항 1

딥러닝 모델을 이용하여 연산을 수행하는 전자 장치에 있어서,

입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하고, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 레이어 관리부;

상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하고, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 성능 확인부;

상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하고, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 지연 제어부;

상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 리소스 관리부; 및

상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 속행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 수신된 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 입출력부를 포함하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 2

제 1 항에 있어서,

상기 레이어 관리부는,

상기 복수의 레이어들 중 상기 입력 데이터에 대한 연산을 수행하는데 이용되는 제 1 복수의 레이어들을 선택하고, 상기 선택된 제 1 복수의 레이어들 각각의 타입을 식별하고, 상기 식별된 제 1 복수의 레이어들 각각의 타입과 대응되는 연산 공식을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 3

제 1 항에 있어서,

미리 설정된 QoS(quality of service) 데이터를 확인하고, 상기 확인된 QoS 데이터에 기초하여 상기 딥러닝 모델에 대한 압축을 수행하는 모델 압축부; 및

상기 압축이 수행된 딥러닝 모델에 대한 프로파일링을 수행하여 상기 압축이 수행된 딥러닝 모델 내 상기 복수의 레이어들 각각에 대한 동작 시간을 측정하고, 상기 측정된 동작 시간을 이용하여 상기 복수의 레이어들 중 미리 설정된 시간 간격마다 동작되는 하나 이상의 레이어들을 선택하고, 상기 선택된 하나 이상의 레이어들 각각에 마일스톤 식별자를 할당하는 마일스톤 식별자 할당부를 포함하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 4

제 1 항에 있어서,

상기 성능 확인부는,

상기 저장된 소요 시간의 상기 미리 설정된 목표 시간에 대한 비율을 상기 지연 값으로 획득하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 5

제 4 항에 있어서,

상기 지연 제어부는,

상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 작으면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 하향 조정하도록 지시하는 제 1 제어 명령을 생성하고,

상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 크면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 상향 조정하도록 지시하는 제 2 제어 명령을 생성하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 6

제 5 항에 있어서,

상기 지연 제어부는,

상기 획득된 지연 값이 상기 임계 범위를 벗어난 정도(degree)를 측정하고, 상기 측정된 정도에 기초하여 상기 적어도 하나의 프로세서의 성능이 하향 조정 또는 상향 조정되는 값을 결정하고, 상기 결정된 값에 따라 상기 제 1 제어 명령 또는 상기 제 2 제어 명령을 생성하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치.

청구항 7

딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법에 있어서,

입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 단계;

상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 단계;

상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하는 단계;

상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 단계;

상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하는 단계;

상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 단계;

상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 단계; 및

상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 속행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 단계를 포함하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 8

제 7 항에 있어서,

상기 입력 데이터에 대한 연산을 수행하는 단계는,

상기 복수의 레이어들 중 상기 입력 데이터에 대한 연산을 수행하는데 이용되는 제 1 복수의 레이어들을 선택하는 단계;

상기 선택된 제 1 복수의 레이어들 각각의 타입을 식별하는 단계; 및

상기 식별된 제 1 복수의 레이어들 각각의 타입과 대응되는 연산 공식을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 단계를 더 포함하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 9

제 7 항에 있어서,

미리 설정된 QoS(quality of service) 데이터를 확인하고, 상기 확인된 QoS 데이터에 기초하여 상기 딥러닝 모델에 대한 압축을 수행하는 단계;

상기 압축이 수행된 딥러닝 모델에 대한 프로파일링을 수행하여 상기 압축이 수행된 딥러닝 모델 내 상기 복수의 레이어들 각각에 대한 동작 시간을 측정하는 단계;

상기 측정된 동작 시간을 이용하여 상기 복수의 레이어들 중 미리 설정된 시간 간격마다 동작되는 하나 이상의 레이어들을 선택하는 단계; 및

상기 선택된 하나 이상의 레이어들 각각에 마일스톤 식별자를 할당하는 단계를 더 포함하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 10

제 7 항에 있어서,

상기 지연 값을 획득하는 단계는,

상기 저장된 소요 시간의 상기 미리 설정된 목표 시간에 대한 비율을 상기 지연 값으로 획득하는 단계를 더 포함하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 11

제 10 항에 있어서,

상기 제어 명령을 생성하는 단계는,

상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 작으면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 하향 조정하도록 지시하는 제 1 제어 명령을 생성하는 단계; 및

상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 크면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 상향 조정하도록 지시하는 제 2 제어 명령을 생성하는 단계를 더 포함하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 12

제 10 항에 있어서,

상기 제어 명령을 생성하는 단계는,

상기 획득된 지연 값이 상기 임계 범위를 벗어난 정도(degree)를 측정하는 단계;

상기 측정된 정도에 기초하여 상기 적어도 하나의 프로세서의 성능이 하향 조정 또는 상향 조정되는 값을 결정하는 단계; 및

상기 결정된 값에 따라 상기 제 1 제어 명령 또는 상기 제 2 제어 명령을 생성하는 단계를 더 포함하는 것을 특징으로 하는, 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법.

청구항 13

제 7 항 내지 제 12 항 중 어느 한 항의 방법을 컴퓨터로 하여금 수행하도록 하는 프로그램을 기록한 컴퓨터 판독 가능 기록 매체.

청구항 14

제 7 항 내지 제 12 항 중 어느 한 항의 방법을 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명은 딥러닝 모델을 통한 추론 서비스를 제공할 때, 적어도 하나의 프로세서의 성능을 제어하는 전자 장치 및 그의 동작 방법에 대한 것이다.

배경 기술

[0003] 딥러닝(deep learning)은 사물이나 데이터를 군집화하거나 분류하는 데 이용되는 기술로서, 다층 구조 형태의 신경망(neural network)을 통해 강건(robust)하고 정확(accurate)한 추론(inference)을 가능하게 한다는 점에서 최근 그 쓰임새가 확장되고 있다. 예를 들어, 딥러닝의 신경망에 의해 구동되는 시각적 장면에 대한 이해와 관련된 기술은 증강 현실용 웨어러블 장치, 홈 오토메이션 장치, 카메라 기반의 감시 장치 및 자율 주행 차량과 같은, 가상 물리 시스템(cyber-physical system)에 활발히 적용되고 있다.

[0004] 이처럼 딥러닝에 이용되는 신경망은 연속적으로 배치된 복수의 레이어들로 구성될 수 있으며, 복수의 레이어들은 배치된 순서에 따라 레이어 별 연산을 수행한 후 연산 결과를 그 다음에 배치된 레이어로 전달하도록 설정될 수 있다.

[0005] 한편, 딥러닝을 통한 추론 과정은 수많은 연산들을 수반하기 때문에, 딥러닝을 이용하여 데이터를 처리하는 전자 장치에는 고성능의 프로세서, 충분한 전력을 공급할 수 있는 전원 장치 및 충분한 저장 공간을 제공할 수 있는 메모리 등이 구비될 필요가 있다. 따라서, 성능, 공급 전력, 저장 공간 등 리소스(resource)가 제한된 모바일 장치에서는 딥러닝의 사용이 제한되는 측면이 있었다.

[0006] 최근에는 모바일 장치에서도 딥러닝에 따른 추론 과정이 원활히 수행될 수 있도록, 리소스를 효율적으로 관리하기 위한 다양한 방법들이 제안되고 있다. 예를 들어, 전자 장치에 하드웨어 기반의 가속기를 적용하는 방법이나, 딥러닝 모델을 압축하여 연산량을 줄이는 방법 등이 제안되고 있으며, 제안된 방법들을 통해 모바일 및 임베디드 장치에서 딥러닝에 따른 추론 과정을 보다 원활하게 수행하는 것이 가능해졌다.

발명의 내용

해결하려는 과제

- [0008] 추론 과정은 딥러닝을 통해 임의의 입력과 대응되는 출력을 획득하는 과정을 의미할 수 있다. 또한, 딥러닝 모델은 추론 과정에 이용될 수 있는 알고리즘으로서, 신경망을 구성하는 복수의 레이어들의 구조를 정의하거나 레이어 별로 수행되는 연산의 매개 변수 또는 가중치를 정의하는데 이용될 수 있다.
- [0009] 일 실시 예에 따르면, 딥러닝 모델에 기초하여 정의된 매개 변수 또는 가중치는 트레이닝 과정을 통해 학습(learning), 즉, 갱신될 수 있다. 트레이닝 과정은 많은 연산 처리량을 필요로 하기 때문에 클라우드 등 서버를 통해 수행될 수 있으며, 이를 통해 학습된 딥러닝 모델은 스마트 폰, 감시 카메라, 스마트 스피커 및 자율 주행 차량 등 다양한 모바일 장치들에 제공될 수 있다. 이렇게 하여, 다양한 모바일 장치들에 제공된 딥러닝 모델은 애플리케이션에서 입력된 데이터를 분류하고 결과를 예측하는 추론 과정에 이용될 수 있다.
- [0010] 한편, 추론 과정은 트레이닝 과정에 비해 상대적으로 적은 연산 처리량을 필요로 할 수 있으나 모바일 장치에서 프로세서의 성능 및 공급 가능한 전력량은 한정적일 수밖에 없으므로, 연산에 따른 레이턴시(latency)와 전력 효율성은 추론 과정에서 주요한 성능 지표로 고려될 수 있다. 예를 들어, 모바일 장치에서 실행되는 STT(speech-to-text) 애플리케이션 또는 번역 애플리케이션의 결과값이 추론 과정을 통해 출력되거나 자율 주행에 이용되는 판단의 결과값이 추론 과정을 통해 출력되는 경우, 각각의 결과값은 최대한 빨리 출력될 것을 요청받을 수 있다. 이 경우, 모바일 장치는 프로세서의 성능을 상향 조정하여 추론 과정에 따른 결과값을 짧은 레이턴시(예: 200 ms)로 출력할 수 있다. 하지만, 이와 동시에 고려되어야 할 사항은 모바일 장치가 공급할 수 있는 전력량으로서, 결과값을 짧은 레이턴시로 출력하기 위해 프로세서의 성능이 상향 조정될 경우 상대적으로 많은 양의 전력이 소모될 수 있기 때문에, 모바일 장치에서 프로세서의 성능을 상향 조정하는 데는 제한이 따를 수밖에 없다. 결국, 모바일 장치에서 추론 과정에 따른 연산이 수행될 때 레이턴시 및 소모 전력은 상호 트레이드 오프(trade off) 관계에 있는 것으로 볼 수 있으며, 이에 따라, 출력되는 결과값 별로 요청되는 최적의 레이턴시를 고려하여 프로세서의 성능이 동적으로 조정될 필요가 있다.
- [0011] 나아가, 모바일 장치에서 추론 과정을 이용하여 판단을 수행함에 있어, 매 판단 시마다 지정된 레이턴시에 따라 일정한 간격으로 결과값이 출력되도록 설정되는 것은 매 판단 시마다 정확한 결과값이 출력되도록 설정되는 것만큼이나 결과값의 신뢰성 제고 측면에서 중요한 요소일 수 있다. 예컨대, 모바일 장치에서 추론 과정을 이용하여 자율 주행에 필요한 판단을 수행할 때, 매 판단 시마다 서로 다른 레이턴시에 따라 결과값이 출력된다면, 결과값이 출력되는 시간을 예측하기 어려워 자율 주행의 안정성이 저해될 수 있으며, 이에 따라, 상기 추론 과정에 이용되는 딥러닝 모델은 자율 주행에 필요한 판단을 수행할 때 이용되기 어려울 수 있다.
- [0012] 본 문서에 개시된 다양한 실시 예들은 딥러닝 모델을 이용하여 추론 과정을 수행함에 있어, 추론 과정에 따른 결과값이 출력되는 시점을 예측 가능한 범위 이내로 설정하기 위하여 프로세서의 성능을 동적으로 조정할 수 있는 방법을 제안한다. 예컨대, 일 실시 예에 따른 전자 장치는 출력되는 결과값 별로 요청되는 최적의 레이턴시를 계산하고, 계산된 레이턴시에 따라 결과값이 출력될 수 있도록 전자 장치 내 적어도 하나의 프로세서의 성능을 조정할 수 있다.

과제의 해결 수단

- [0014] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치는, 입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하고, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 레이어 관리부, 상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하고, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 성능 확인부, 상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하고, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 지연 제어부, 상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 리소스 관리부, 및 상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 속행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 수신된 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 입출력부를 포함할 수 있다.

[0015] 또한, 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법은, 입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 단계, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 단계, 상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하는 단계, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 단계, 상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하는 단계, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 단계, 상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 단계, 및 상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 속행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 단계를 포함할 수 있다.

발명의 효과

[0017] 본 문서에 개시된 다양한 실시 예들에 따르면, 전자 장치는 딥러닝 모델을 통해 추론 과정을 수행함에 있어 추론 과정에 따른 결과값이 예측 가능한 시간 범위 안에서 출력되도록 설정될 수 있다. 이처럼, 추론 과정에 따른 결과값이 예측 가능한 시간 범위 안에서 출력되도록 설정됨에 따라, 딥러닝 모델을 이용하는 전자 장치의 안정성과 딥러닝에 따른 연산의 신뢰성이 보장될 수 있다.

[0018] 또한, 본 문서에 개시된 다양한 실시 예들에 따르면, 전자 장치는 추론 과정에 따른 결과값이 지정된 레이턴시에 따라 출력되도록 하기 위해 적어도 하나의 프로세서의 성능을 동적으로 조정할 수 있다. 이에 따라, 프로세서의 성능이 필요 이상으로 상향 조정되는 경우가 사라지게 됨에 따라, 전자 장치 내 소모 전력을 효율적으로 관리할 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일 실시 예에 따른 전자 장치의 구성을 도시한 도면이다.

도 2는 본 발명의 일 실시 예에 따른 전자 장치에 구비되는 런타임 모듈의 구성을 도시한 도면이다.

도 3은 본 발명의 일 실시 예에 따른 전자 장치에 구비되는 피드백 회로의 구성을 도시한 도면이다.

도 4는 본 발명의 일 실시 예에 따라 딥러닝 모델에 마일스톤 식별자를 할당하는 방법을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시 예에 따른 전자 장치에서 딥러닝 모델에 마일스톤 식별자를 할당하는 동작을 설명하기 위한 순서도이다.

도 6은 본 발명의 일 실시 예에 따른 전자 장치에서 딥러닝 모델을 통한 추론 과정을 수행할 때, 적어도 하나의 프로세서의 성능을 제어하는 방법을 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 문서에 개시된 다양한 실시 예들은 본 발명을 특정한 실시 형태로 한정하기 위해 제시된 것이 아니며, 다양한 실시 예들을 통해 소개된 구성요소들은 본 발명의 사상 및 기술 범위에 포함되는 모든 변경 가능한 균등물 내지 대체물을 포함하는 의미로서 제시된 것임을 당업자는 용이하게 이해할 것이다. 또한, 각 도면을 설명함에 있어, 다르게 정의되지 않는 한 기술적이거나 과학적인 용어를 포함해서 본 명세서 상에서 사용되는 모든 용어들은 본 발명이 속하는 기술분야에서 통상의 지식을 가진 사람에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있는 것으로 해석될 수 있다. 또한, 본 발명의 목적 및 효과, 그리고 그것들을 달성하기 위한 기술적 구성들은 첨부되는 도면과 함께 상세하게 설명되는 실시 예들을 통해 명확해질 것이다. 본 발명을 설명함에 있어 공지 기능 또는 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우, 그와 관련된 상세한 설명은 생략될 수 있으며, 뒤에 설명되는 용어들은 본 발명에서의 구조, 역할 및 기능 등을 고려하여 정의된 용어들로서 이는 사용자 및 운용자의 의도 또는 관례 등에 따라 기존에 사용되던 의미와 달리

해석될 수 있다.

[0022] 본 발명은 이하에서 개시되는 실시 예들에 한정되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있음을 밝혀둔다. 본 문서에 개시된 다양한 실시 예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 오로지 특허청구범위에 기재된 청구항의 범주에 의하여 정의될 뿐이다.

[0023] 본 문서에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있다는 것을 의미한다. 또한, 본 발명의 다양한 실시 예들에 있어서, 각 구성요소들, 기능 블록들 또는 수단들은 하나 또는 그 이상의 하부 구성요소로 구성될 수 있고, 각 구성요소들이 수행하는 전기, 전자, 기계적 기능들은 전자회로, 집적회로, ASIC(application specific integrated circuit) 등 공지된 다양한 소자들 또는 기계적 요소들로 구현될 수 있으며, 각각 별개로 구현되거나 2 이상이 하나로 통합되어 구현될 수도 있다.

[0024] 한편, 첨부된 블록도의 블록들이나 흐름도의 단계들은 범용 컴퓨터, 특수용 컴퓨터, 휴대용 노트북 컴퓨터, 네트워크 컴퓨터 등 데이터 프로세싱이 가능한 장비의 프로세서나 메모리에 탑재되어 지정된 기능들을 수행하는 컴퓨터 프로그램 인스트럭션들을 의미하는 것으로 해석될 수 있다. 이들 컴퓨터 프로그램 인스트럭션들은 컴퓨터 장치에 구비된 메모리 또는 컴퓨터에서 판독 가능한 메모리에 저장될 수 있기 때문에, 블록도의 블록들 또는 흐름도의 단계들에서 설명된 기능들은 이를 수행하는 인스트럭션 수단을 내포하는 제조물로 생산될 수도 있다. 아울러, 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 가능한 실시 예들에서는 블록들 또는 단계들에서 언급된 기능들이 정해진 순서와 달리 실행되는 것도 가능함을 주목해야 한다. 예컨대, 잇달아 도시되어 있는 두 개의 블록들 또는 단계들은 실질적으로 동시에 수행되거나, 역순으로 수행될 수 있으며, 경우에 따라 일부 블록들 또는 단계들이 생략된 채로 수행될 수도 있다.

[0026] 도 1은 본 발명의 일 실시 예에 따른 전자 장치의 구성을 도시한 도면이다. 다양한 실시 예들에 따르면, 전자 장치(100)는 애플리케이션(110), 런타임 모듈(120), 프로세서(130), 센서(140), 저장부(150) 및 입출력부(160) 중 적어도 하나를 포함할 수 있다. 여기서, 런타임 모듈(120)은 특정 인스트럭션을 수행하는 모듈을 의미하기 위해 사용되었으며, 상기 모듈들을 통해 수행되는 인스트럭션들은 전자 장치의 프로세서(130)에 의해 수행되는 것으로 이해될 수 있음을 밝혀둔다.

[0027] 전자 장치(100)의 애플리케이션(110)은 특정 업무를 수행할 수 있는 프로그램 또는 상기 프로그램이 실행되는 모듈을 의미할 수 있다. 일 실시 예에 따르면, 애플리케이션(110)은 딥러닝 모델을 통한 추론 과정을 호출하거나 추론 과정을 통해 획득된 결과 데이터를 미리 정해진 방식에 따라 사용자에게 제공하는 동작을 수행할 수 있다. 애플리케이션(110)은 입력 데이터가 수신되는 것에 응답하여 추론 과정을 호출하도록 설정될 수 있으며, 호출된 추론 과정은 센서(140)에서 수신되는 데이터에 기초하여 런타임 모듈(120)을 통해 주기적 또는 비주기적으로 수행될 수 있다.

[0028] 일 실시 예에 따르면, 애플리케이션(110)은 미리 설정된 QoS(quality of service) 데이터를 참고하여 특정 딥러닝 모델(M)을 호출할 수 있다. 여기서, QoS 데이터란 서비스 품질을 나타내는 지표로서, 상기 지표는 네트워크 또는 프로세서 상에서 일정 정도 이하의 레이턴시나 데이터 손실률 등을 보장하기 위한 용도로 이용될 수 있다. 예컨대, QoS 데이터를 나타내는 Q는 하기 수학적 식 1과 같이 정의될 수 있다.

수학적 식 1

[0029] $Q = \leq d, C >$

[0030] 상기 수학적 식 1에서, d는 딥러닝 모델(M)을 통해 추론 과정을 수행함에 있어, 출력 데이터의 획득이 요청되는 제 1 응답 시간을 나타내는 값일 수 있다. 예컨대, 제 1 응답 시간은 추론 과정이 시작되는 시점부터 종료되는 시점까지의 소요 시간을 의미할 수 있다. 또한, 상기 수학적 식 1에서, C는 딥러닝 모델(M)에 대한 압축 범위(compression bound)를 나타내는 값일 수 있다.

- [0031] 일 실시 예에 따르면, 애플리케이션(110)은 미리 설정된 QoS 데이터에 기초하여 딥러닝 모델(M)이 호출되는 경우, QoS 데이터와 함께 딥러닝 모델(M)을 런타임 모듈(120)로 전달할 수 있다.
- [0032] 전자 장치(100)의 런타임 모듈(120)은 입출력부(160)를 통해 수신되는 입력 데이터와 애플리케이션(110)으로부터 수신되는 딥러닝 모델을 이용하여, 딥러닝 연산을 수행할 수 있다. 예컨대, 런타임 모듈(120)은 입력 데이터로서 "고양이 사진"이 수신되고 입력 데이터가 무엇인지 묻는 질문 데이터가 식별되는 경우, 식별된 질문 데이터에 대한 답변 데이터를 생성하기 위해 딥러닝 모델에 포함된 복수의 레이어들 중 하나 이상의 레이어들 각각에 정의된 연산 방법을 이용하여 입력 데이터에 대한 연산을 수행할 수 있다. 런타임 모듈(120)은 연산을 수행한 결과로서 "고양이"라는 답변 데이터가 획득할 수 있으며, 획득된 답변 데이터를 입력 데이터의 추론 결과로서 애플리케이션(110)에 전달할 수 있다.
- [0033] 일 실시 예에 따르면, 런타임 모듈(120)은 애플리케이션(110)으로부터 QoS 데이터 및 딥러닝 모델이 수신되는 경우, QoS 데이터를 확인하여 딥러닝 모델에 대한 압축을 수행할 수 있다. 예를 들어, 런타임 모듈(120)은 QoS 데이터로부터 출력 데이터의 획득이 요청되는 제 1 응답 시간을 확인할 수 있다. 또한, 런타임 모듈(120)은 딥러닝 모델 내 복수의 레이어들 각각을 이용하여 연산을 수행하는데 소요되는 시간을 합산하여 총 소요 시간을 확인할 수 있다. 만약, 총 소요 시간이 제 1 응답 시간을 초과하는 것으로 판단되는 경우, 런타임 모듈(120)은 딥러닝 모델에 대한 압축을 수행하여 딥러닝 모델을 이용한 연산에 소요되는 시간을 줄일 수 있다.
- [0034] 일 실시 예에 따르면, 런타임 모듈(120)은 딥러닝 모델을 이용하여 입력 데이터에 대한 연산을 수행하면서 연산이 진행되는 속도를 일정한 간격으로 확인할 수 있으며, 확인 결과에 기초하여 프로세서(130)의 동작 속도를 제어함으로써 연산에 따른 결과 데이터가 목표 시간보다 빨리 또는 늦게 출력되지 않도록 할 수 있다. 예컨대, 런타임 모듈(120)은 딥러닝 모델에 대한 압축을 수행할 때, 딥러닝 모델 내 복수의 레이어들 중 적어도 하나의 레이어에 제 1 마일스톤 식별자를 할당할 수 있다. 이 후 런타임 모듈(120)은 딥러닝 모델을 이용하여 입력 데이터에 대한 연산을 수행하던 중 미리 할당된 제 1 마일스톤 식별자가 확인되면, 연산이 시작된 시점 또는 앞서 다른 마일스톤 식별자가 확인된 시점부터 제 1 마일스톤 식별자가 확인된 시점까지의 연산에 실제로 소요된 시간인 제 2 응답 시간을 측정할 수 있다. 이와 동시에, 런타임 모듈(120)은 연산이 시작된 시점 또는 앞서 다른 마일스톤 식별자가 확인된 시점부터 제 1 마일스톤 식별자가 확인된 시점까지의 연산에 소요되어야 하는 제 1 목표 시간을 확인할 수 있다. 다양한 실시 예들에 따르면, 런타임 모듈(120)은 제 k 시점에 제 2 응답 시간의 제 1 목표 시간에 대한 비율을 지연 값으로서 획득할 수 있으며, 제 k 시점의 지연 값은 하기 수학적 식 2와 같이 정의될 수 있다.

수학적 식 2

$$\text{지연 값}(k) = \frac{\text{실제 응답 시간}(k)}{\text{목표 응답 시간}(k)}$$

- [0035]
- [0036] 만약, 지연 값이 1보다 크다면, 런타임 모듈(120)을 통한 연산은 목표로 설정된 시간보다 느리게 수행되고 있는 것으로 판단될 수 있다. 반면에, 지연 값이 1보다 작다면, 런타임 모듈(120)을 통한 연산은 목표로 설정된 속도보다 빠르게 수행되고 있는 것으로 판단될 수 있다.
- [0037] 일 실시 예에 따르면, 런타임 모듈(120)은 연산이 목표로 설정된 속도보다 느리게 수행되고 있는 것으로 판단되는 경우, 연산에 이용되는 프로세서(130)의 동작 속도를 상향 조정할 수 있다. 또한, 런타임 모듈(120)은 연산이 목표로 설정된 시간보다 빠르게 수행되고 있는 것으로 판단되는 경우, 연산에 이용되는 프로세서(130)의 동작 속도를 하향 조정할 수 있다.
- [0038] 전자 장치(100)의 프로세서(130)는 런타임 모듈(120)에서 수행되는 연산 중 적어도 일부를 처리할 수 있다. 예컨대, 프로세서(130)는 딥러닝 모델 내 복수의 레이어들 중 적어도 하나의 레이어에서 수행되는 연산을 수행한 후, 그 결과 데이터를 런타임 모듈(120)로 제공할 수 있다. 다양한 실시 예들에 따르면, 프로세서(130)는 런타임 모듈(120)을 포함하는 형태로 구현될 수 있으며, 이 경우 런타임 모듈(120)에서 수행되는 모든 인스트럭션들은 프로세서(130)에서 수행되는 것으로 이해될 수 있다.
- [0039] 일 실시 예에 따르면, 프로세서(130)는 하나 이상의 프로세서들을 포함할 수 있다. 예컨대, 프로세서(130)는 하나 이상의 중앙처리장치(central processing unit, CPU)들 및 하나 이상의 그래픽처리장치(graphics

processing unit, GPU)들을 포함할 수 있다. 프로세서(130)는 딥러닝 모델 내 복수의 레이어들 각각의 타입에 기초하여 연산을 수행하는데 필요한 처리장치를 선택할 수 있으며, 선택된 처리장치를 이용하여 연산을 수행할 수 있다.

[0040] 일 실시 예에 따르면, 프로세서(130)는 런타임 모듈(120)로부터 동작 속도를 상향 조정하는 것과 대응되는 제어 명령이 수신되는 경우, 수신된 제어 명령과 대응되는 프로세서(130)의 주파수를 상향 조정하여 연산 처리 속도를 높일 수 있다. 반면에, 프로세서(130)는 런타임 모듈(120)로부터 동작 속도를 하향 조정하는 것과 대응되는 제어 명령이 수신되는 경우, 수신된 제어 명령과 대응되는 프로세서(130)의 주파수를 하향 조정하여 연산 처리 속도를 낮출 수 있다.

[0041] 전자 장치(100)의 센서(140)는 전자 장치(100)의 내부의 작동 상태(예: 전력 또는 온도), 또는 외부의 환경 상태에 대응하는 전기 신호 또는 데이터 값을 생성할 수 있다. 예컨대, 센서(140)는 제스처 센서, 자이로 센서, 기압 센서, 마그네틱 센서, 가속도 센서, 그립 센서, 근접 센서, 컬러 센서, IR(infrared) 센서, 생체 센서, 온도 센서, 습도 센서 및 조도 센서 중 적어도 하나를 포함할 수 있다. 일 실시 예에 따르면, 전자 장치(100)는 자율 주행 차량에 장착된 후 센서(140)로부터 획득되는 입력 데이터를 이용하여 딥러닝 모델을 통한 연산을 수행할 수 있고, 딥러닝 모델을 통한 연산의 결과 값으로 출력 데이터를 생성하여 자율 주행 차량에 제공할 수 있다.

[0042] 전자 장치(100)의 저장부(150)는 전자 장치(100)의 적어도 하나의 구성 요소에 의해 사용되는 다양한 데이터를 저장할 수 있다. 여기서, 데이터는 소프트웨어 및 이와 관련된 명령에 대한 입력 데이터 또는 출력 데이터를 포함할 수 있다. 또한, 저장부(150)는 휘발성 메모리 또는 비휘발성 메모리를 포함할 수 있다. 일 실시 예에 따르면, 저장부(150)는 입력 데이터에 대한 연산을 수행할 때 이용될 수 있는 하나 이상의 딥러닝 모델들과 하나 이상의 딥러닝 모델들에 포함된 복수의 레이어들에 대한 정보를 저장할 수 있다.

[0043] 전자 장치(100)의 입출력부(160)는 전자 장치(100)의 구성 요소에 사용될 명령 또는 데이터를 전자 장치(100)의 외부(예: 사용자)로부터 수신하거나 외부로 제공할 수 있다. 입출력부(160)는 입력부 및 출력부로 구분될 수 있으며, 입력부는 마우스, 키보드 및 터치 패드를 포함할 수 있고 출력부는 디스플레이 및 스피커를 포함할 수 있다.

[0045] 도 2는 본 발명의 일 실시 예에 따른 전자 장치에 구비되는 런타임 모듈의 구성을 도시한 도면이다. 다양한 실시 예들에 따르면, 전자 장치(100)는 런타임 모듈(120), 프로세서(130), 저장부(150) 및 입출력부(160) 중 적어도 하나를 포함할 수 있다. 또한, 런타임 모듈(120)은 QoS 관리부(200) 및 실행 관리부(210)를 포함할 수 있다. 또한, QoS 관리부(200)는 모델 압축부(201), 마일스톤 식별자 할당부(203), 성능 확인부(205), 지연 제어부(207) 및 리소스 관리부(209)를 포함할 수 있다. 또한, 실행 관리부(210)는 데이터 전처리부(211), 모델 관리부(213) 및 레이어 관리부(215)를 포함할 수 있다. 또한, 프로세서(130)는 중앙처리장치(220) 및 그래픽처리장치(225)를 포함할 수 있다.

[0046] 일 실시 예에 따르면, QoS 관리부(200)는 QoS 데이터에 기초하여, 딥러닝 모델을 압축하거나 프로세서(130)의 동작 속도를 제어할 수 있다. 실행 관리부(210)는 딥러닝 모델을 통해 입력 데이터에 대한 연산을 수행하거나 딥러닝 모델 내 적어도 일부의 레이어들에 할당된 마일스톤 식별자를 식별할 수 있다.

[0047] QoS 관리부(200)의 모델 압축부(201)는 미리 설정된 QoS 데이터를 통해 확인된 d (딥러닝 모델을 통해 추론 과정을 수행함에 있어, 출력 데이터의 획득이 요청되는 제 1 응답 시간) 및 C (딥러닝 모델에 대한 압축 범위)를 이용하여, 딥러닝 모델에 대한 압축이 필요한 지 여부를 판단할 수 있다. 예컨대, 추론 과정에 대한 호출이 있을 경우, 모델 압축부(201)는 상기 호출에 따라 선택된 딥러닝 모델의 메모리 사용량 및 전자 장치(100)의 가용 메모리의 양을 체크할 수 있으며, 상기 체크 결과에 따라 딥러닝 모델에 대한 압축량을 결정할 수 있다. 한편, 모델 압축부(201)는 딥러닝 모델에 대한 압축을 수행함에 있어, 추론 과정의 정확도가 예기치 않을 정도로 손실되는 것을 방지하기 위하여, 단계적인 압축 방법을 이용할 수 있다. 예컨대, 모델 압축부(201)는 딥러닝 모델에 대한 압축을 할 때, 적절한 순위를 선택함으로써 압축의 수준을 제어할 수 있는 특이값 분해 (singular value decomposition) 기술을 통한 근사 방법을 이용할 수 있다. 또는, 모델 압축부(201)는 딥러닝 모델을 통한 연산에 소요되는 총 소요 시간 및 제 1 응답 시간 간의 비교를 통해 딥러닝 모델에 대한 압축의 수준을 결정할 수 있다. 예컨대, 딥러닝 모델을 통한 연산에 소요되는 총 소요 시간이 제 1 응답 시간보다 긴 것으로 판단되는 경우, 모델 압축부(201)는 총 소요 시간 및 제 1 응답 시간 간의 차이를 계산할 수 있으며 계산 결과에 따라 상기

딤러닝 모델에 대한 압축 수준을 결정할 수 있다. 한편, 모델 압축부(201)는 딤러닝 모델에 대한 압축 수준이 결정되면 이와 대응되는 제어 명령을 생성하고, 생성된 제어 명령을 모델 관리부(213)로 전달할 수 있다.

[0048] QoS 관리부(200)의 마일스톤 식별자 할당부(203)는 압축이 수행된 딤러닝 모델에 대하여 실행 상태나 통신 상태 등을 해석하는 프로파일링을 수행할 수 있다. 일 실시 예에 따르면, 마일스톤 식별자 할당부(203)는 압축이 수행된 딤러닝 모델 내 복수의 레이어들 각각에 대한 동작 시간을 측정할 수 있다. 이어서, 마일스톤 식별자 할당부(203)는 측정된 동작 시간을 이용하여, 목표로 설정된 시간 간격마다 위치한 레이어들에 마일스톤 식별자를 할당하기 위한 제어 명령을 생성할 수 있다. 예컨대, 마일스톤 식별자 할당부(203)는 압축이 수행된 딤러닝 모델을 통한 연산이 시작된 시점부터 100 ms 간격으로 동작되는 레이어들을 식별할 수 있고, 식별된 레이어들 각각에 마일스톤 식별자를 할당하기 위한 제어 명령을 생성할 수 있다. 보다 구체적으로, 마일스톤 식별자 할당부(203)는 연산이 시작된 후 100 ms 이 경과한 시점에 동작하는 제 1 레이어에 제 1 마일스톤 식별자를 할당하고, 이후 또 다시 100 ms 이 경과한 시점에 동작하는 제 2 레이어에 제 2 마일스톤 식별자를 할당하기 위한 제어 명령을 생성할 수 있다. 한편, 마일스톤 식별자 할당부(203)는 마일스톤 식별자를 할당하기 위해 생성된 제어 명령을 모델 관리부(213)로 전달할 수 있다.

[0049] 실행 관리부(210)의 데이터 전처리부(211)는 압축이 수행된 딤러닝 모델에 마일스톤 식별자를 할당하는 동작이 완료된 후, 입력 데이터가 외부로부터 수신되는 것에 응답하여, 딤러닝 모델을 통한 연산에 이용될 수 있도록 수신된 입력 데이터의 형식 또는 내용을 수정할 수 있다.

[0050] 실행 관리부(210)의 모델 관리부(213)는 모델 압축부(201)로부터 수신된 제어 명령 또는 마일스톤 식별자 할당부(203)로부터 수신된 제어 명령에 기초하여, 딤러닝 모델에 대한 압축을 수행하거나 딤러닝 모델 내 적어도 하나의 레이어에 마일스톤 식별자를 할당할 수 있다.

[0051] 실행 관리부(210)의 레이어 관리부(215)는 데이터 전처리부(211)로부터 형식 또는 내용이 수정된 입력 데이터가 수신되면, 상기 압축이 수행된 딤러닝 모델과 상기 수정된 입력 데이터를 이용하여 추론 과정에 따른 연산을 시작할 수 있다. 예컨대, 레이어 관리부(215)는 압축이 수행된 딤러닝 모델 내 복수의 레이어들 각각의 타입과 대응되는 연산 공식을 이용하여 연산을 진행할 수 있다. 이 때, 연산을 진행하는 순서는 딤러닝 모델 내 복수의 레이어들이 배치된 순서에 따라 진행될 수 있다.

[0052] 일 실시 예에 따르면, 레이어 관리부(215)는 연산을 진행 중인 레이어에 마일스톤 식별자가 할당되어 있는지 여부를 확인할 수 있다. 만약, 미리 할당된 제 1 마일스톤 식별자가 확인된다면, 레이어 관리부(215)는 연산이 시작된 시점부터 제 1 마일스톤 식별자가 확인된 시점까지 소요된 시간을 측정한 후, 측정된 소요 시간에 대한 제 1 데이터를 QoS 관리부(200)의 성능 확인부(205)로 전달할 수 있다. 이 후, 레이어 관리부(215)는 연산을 속행하여 미리 할당된 제 2 마일스톤 식별자가 확인된다면, 제 1 마일스톤 식별자가 확인된 시점부터 제 2 마일스톤 식별자가 확인된 시점까지 소요된 시간을 측정한 후, 측정된 소요 시간에 대한 제 2 데이터를 앞서와 마찬가지로 QoS 관리부(200)의 성능 확인부(205)로 전달할 수 있다. 레이어 관리부(215)는 종결 식별자가 확인되기 전까지 연산-식별자 확인-소요 시간 측정-측정 값 전달을 반복할 수 있고, 종결 식별자가 확인되는 것에 응답하여 연산을 종료할 수 있다. 레이어 관리부(215)는 연산이 종료되는 경우, 입력 데이터와 대응되는 출력 데이터로서 연산이 종료되는 시점까지 연산된 데이터를 입출력부(160)를 통해 사용자에게 제공할 수 있다.

[0053] QoS 관리부(200)의 성능 확인부(205)는 레이어 관리부(215)로부터 마일스톤 식별자가 확인되는 시점까지의 소요 시간에 대한 데이터가 수신되는 경우, 수신된 데이터를 이용하여 전자 장치(100) 내 적어도 하나의 구성 요소에 대한 성능을 확인할 수 있다. 일 실시 예에 따르면, 미리 할당된 제 1 마일스톤 식별자가 확인되는 것에 응답하여 레이어 관리부(215)로부터 연산이 시작된 시점부터 제 1 마일스톤 식별자가 확인된 시점까지의 소요 시간에 대한 제 1 데이터가 수신되는 경우, 성능 확인부(205)는 수신된 제 1 데이터에 따른 소요 시간과 미리 설정된 목표 시간을 비교하여 지연 값을 획득할 수 있다. 이 때, 지연 값은 수학적 2에 따라 획득될 수 있다. 예컨대, 제 1 데이터에 따른 소요 시간이 120 ms 으로 측정되고, 목표 시간이 100 ms 으로 설정된 경우, 지연 값은 $1.2(=120 \text{ ms} / 100 \text{ ms})$ 로 계산될 수 있다. 한편, 이와 같은 방식으로 계산된 지연 값은 성능 확인부(205)에서 지연 제어부(207)로 전달될 수 있다.

[0054] QoS 관리부(200)의 지연 제어부(207)는 성능 확인부(205)로부터 수신된 지연 값에 기초하여, 실제 연산이 목표로 설정된 속도보다 빠르게 진행되고 있는지 또는 느리게 진행되고 있는지 여부를 판단할 수 있다. 예컨대, 지연 제어부(207)는 지연 값으로 1.2가 수신되는 경우(즉, 지연 값이 1보다 큰 경우), 실제 연산이 목표로 설정된 속도보다 느리게 진행되고 있는 것으로 판단할 수 있다. 반면에, 지연 제어부(207)는 지연 값이 1보다 작은 경우, 실제 연산이 목표로 설정된 속도보다 빠르게 진행되고 있는 것으로 판단할 수 있다.

[0055] 일 실시 예에 따르면, 지연 제어부(207)는 실제 연산이 목표로 설정된 속도보다 느리게 진행되고 있는 것으로 판단되면(지연 값이 1보다 큰 경우), 프로세서(130)의 성능을 상향 조정하도록 지시하는 제어 명령을 생성할 수 있다. 반면에, 지연 제어부(207)는 실제 연산이 목표로 설정된 속도보다 빠르게 진행되고 있는 것으로 판단되면(지연 값이 1보다 작은 경우), 프로세서(130)의 성능을 상향 조정하도록 지시하는 제어 명령을 생성할 수 있다. 한편, 지연 제어부(207)는 프로세서의 성능을 상향 조정 또는 하향 조정하도록 지시하는 제어 명령을 생성함에 있어, 하기 표 1과 같이 미리 설정된 임계 범위를 참고할 수 있다.

표 1

	$0 < 1 - \text{지연 값} \leq 0.1$	$0.1 < 1 - \text{지연 값} \leq 0.3$	$0.3 < 1 - \text{지연 값} $
상향 조정	제 1 레벨 상향 조정	제 2 레벨 상향 조정	제 3 레벨 상향 조정
하향 조정	제 1 레벨 하향 조정	제 2 레벨 하향 조정	제 3 레벨 하향 조정

[0057] 상기 표 1에서, 제 1 레벨, 제 2 레벨 및 제 3 레벨은 프로세서의 성능을 조정하는 정도(degree)와 관련된 값으로서, 사용자에게 의해 미리 설정될 수 있다. 예컨대, 제 1 레벨 상향 조정(또는 하향 조정)은 프로세서의 주파수를 5%만큼 상향 조정(또는 하향 조정)하는 것을 의미할 수 있다. 또는, 제 1 레벨 상향 조정(또는 하향 조정)은 프로세서의 주파수를 50MHz만큼 상향 조정(또는 하향 조정)하는 것을 의미할 수 있다. 예컨대, 앞선 실시 예와 같이, 지연 값이 1.2로 측정되는 경우, 지연 값은 1보다 크므로 상향 조정이 필요한 경우에 해당하고, 1과 지연 값의 차이는 0.2로 계산될 수 있다. 이 경우, 지연 제어부(207)는 상기 표 1에 정의된 바를 참고하여, 프로세서의 성능이 제 2 레벨 상향 조정될 필요가 있는 것으로 판단할 수 있고, 이와 대응되는 제어 명령을 생성할 수 있다.

[0058] 일 실시 예에 따르면, 지연 제어부(207)는 프로세서(130)의 성능을 상향 조정(또는 하향 조정)하도록 지시하는 제어 명령을 생성함에 있어, PI(proportional integral) 제어를 통해 근사화된 하기 수학식 3을 이용할 수 있다.

수학식 3

[0059]
$$\text{speed}(k) = (K_p + K_I) \times e(k) - K_p \times e(k-1)$$

[0060] 상기 수학식 3에서, speed(k)는 k 시점 이전에 획득된 데이터에 기초하여 k 시점에 필요한 프로세서(130)의 성능을 의미할 수 있다. 예컨대, k-1시점의 프로세서 성능 speed(k-1)보다 k시점의 프로세서의 성능 speed(k)가 10% 큰 경우, 프로세서의 주파수를 10% 증가시켜야 한다. 반대로, speed(k)가 speed(k-1)보다 적은 경우, 프로세서의 주파수를 감소시켜야 한다. 여기서, K_p 및 K_I 는 실험적으로 획득된 PI 제어기의 이득이고, e(k)는 에러 값으로서 k 시점에 획득된 지연 값과 기준 값의 차이를 의미할 수 있다. 예컨대, k 시점에 획득된 지연 값을 지연값(k)라고 정의하고, 기준 값을 1로 정의하면, e(k)는 1-지연값(k)를 의미할 수 있다.

[0061] 다양한 실시 예들에 따르면, 상기 수학식 3의 K_p 및 K_I 는 하기 표 2와 같이 실험적으로 획득될 수 있다.

표 2

	K_p	K_I
CPU	-0.28	-0.42
GPU	-0.60	-0.60

[0063] 상기 표 2에 따르면, 조정이 필요한 프로세서(130)가 중앙처리장치(CPU)인지 그래픽처리장치(GPU)인지에 따라 서로 다른 값들이 획득될 수 있다.

[0064] QoS 관리부(200)의 리소스 관리부(209)는 지연 제어부(207)에서 생성된 제어 명령에 따라, 전자 장치(100) 내 적어도 하나의 구성 요소의 성능을 제어할 수 있다. 예컨대, 리소스 관리부(209)는 지연 제어부(207)로부터 수

신된 제어 명령에 기초하여, 프로세서(130)의 성능을 조정할 수 있다. 이 때, 리소스 관리부(209)는 DVFS(dynamic voltage / frequency scaling)와 같은 튜닝 노브(tuning knob)를 이용하여 프로세서의 성능을 조정할 수 있다.

[0065] 프로세서(130)는 리소스 관리부(209)의 제어에 따라, 프로세서(130)에 포함된 중앙처리장치(220) 또는 그래픽처리장치(225)의 성능(예: 주파수)을 조정할 수 있다. 성능이 조정된 프로세서(130)는 레이어 관리부(215)에서 딥러닝 모델에 따른 연산이 수행될 때, 조정 전과 다른 연산 속도를 제공할 수 있다.

[0067] 도 3은 본 발명의 일 실시 예에 따른 전자 장치에 구비되는 피드백 회로의 구성을 도시한 도면이다.

[0068] 일 실시 예에 따르면, 제 k 시점에, 마일스톤 식별자가 식별되는 경우, 성능 확인부(205)는 지연 값인 $tard(k)$ 를 획득할 수 있다. 이어서, 지연 제어부(207)는 지연 값과 기준 값으로 설정된 1의 차이인 $e(k)$ 를 획득할 수 있다. 지연 제어부(207)는 획득된 $e(k)$ 에 기초하여 프로세서(130)의 성능이 조정될 필요가 있는지 여부와, 조정이 필요하다면 어느 정도 조정될 필요가 있는지 여부를 판단할 수 있다. 예컨대, 지연 값이 1.2 인 경우, 지연 제어부(207)는 지연 값이 1보다 크므로 프로세서의 성능이 상향 조정될 필요가 있다고 판단할 수 있다. 이 때, $e(k)$ 는 0.2로 확인되는바, 지연 제어부(207)는 최종적으로 프로세서의 성능이 제 2 레벨 상향 조정될 필요가 있는 것으로 결정할 수 있다.

[0069] 지연 제어부(207)는 프로세서(130)의 성능을 제 2 레벨 상향 조정하기 위한 제어 명령으로서 $\Delta speed(k+1)$ 을 생성할 수 있고, 생성된 제어 명령을 DVFS 관리부(300)로 전달할 수 있다.

[0070] 또한, DVFS 관리부(300)는 지연 제어부(207)로부터 수신된 제어 명령에 대응하여, 프로세서(130) 내 중앙처리장치(220) 및 그래픽처리장치(225) 중 적어도 하나의 주파수를 조정하기 위한 제어 명령으로서 $freq(k+1)$ 을 생성할 수 있고, 생성된 제어 명령을 중앙처리장치(220) 및 그래픽처리장치(225) 중 적어도 하나로 전달할 수 있다. 이 후, 조정된 성능에 따라 딥러닝 모델을 통한 연산이 속행될 수 있다.

[0071] 일 실시 예에 따라, 제 k+1 시점에, 마일스톤 식별자가 식별되는 경우, 성능 확인부(205)는 지연 값인 $tard(k+1)$ 를 획득할 수 있다. 이어서, 지연 제어부(207)는 지연 값과 기준 값으로 설정된 1의 차이인 $e(k+1)$ 를 획득할 수 있으며, 프로세서(130)의 성능을 조정하는 동작을 반복하여 수행할 수 있다.

[0073] 도 4는 본 발명의 일 실시 예에 따라 딥러닝 모델에 마일스톤 식별자를 할당하는 방법을 설명하기 위한 도면이다.

[0074] 전자 장치(100)는 "고양이 사진"이 입력 데이터(400)로 수신되는 경우, 딥러닝 모델(420)을 이용하여 입력 데이터(400)에 대한 연산을 수행할 수 있으며, 연산의 결과 데이터(410)로서 "Cat!"을 획득할 수 있다. 한편, 딥러닝 모델(420)은 타입에 따라 서로 다른 연산을 수행하는 복수의 레이어들로 구성될 수 있고, 복수의 레이어들은 미리 설정된 순서에 따라 나열된 형태로 도식화 될 수 있다.

[0075] 일 실시 예에 따르면, 마일스톤 식별자 할당부(203)는 딥러닝 모델을 통한 연산을 수행하기에 앞서, 딥러닝 모델을 통한 연산에 소요되는 총 소요 시간을 측정할 수 있다. 예컨대, 마일스톤 식별자 할당부(203)는 딥러닝 모델을 통한 연산에 소요되는 총 소요 시간이 300 ms 인 것으로 확인할 수 있다. 이 때, 마일스톤 식별자 할당부(203)는 딥러닝 모델 내 복수의 레이어들 중 마지막 레이어에 종료 식별자를 할당할 수 있다.

[0076] 또한, 마일스톤 식별자 할당부(203)는 총 소요 시간을 기준으로 하여, 미리 설정된 시간 간격마다 마일스톤 식별자를 할당할 수 있다. 예컨대, 마일스톤 식별자 할당부(203)는 총 소요 시간의 1/3과 대응되는 시점마다 위치한 레이어에 마일스톤 식별자를 할당할 수 있다. 즉, 마일스톤 식별자 할당부(203)는 총 소요 시간의 1/3과 대응되는 시점에 위치한 제 1 레이어(430) 및 2/3과 대응되는 시점에 위치한 제 2 레이어(440) 각각에 제 1 마일스톤 식별자 및 제 2 마일스톤 식별자를 할당할 수 있다. 만약, 총 소요 시간이 300 ms 라면, 제 1 마일스톤 식별자는 연산이 시작된 시점으로부터 100 ms 이 경과한 시점에 연산을 수행하는 제 1 레이어(430)에 할당될 수 있고, 제 2 마일스톤 식별자는 제 1 마일스톤 식별자가 식별된 시점으로부터 100 ms 이 경과한 시점에 연산을 수행하는 제 2 레이어(440)에 할당될 수 있다.

- [0078] 도 5는 본 발명의 일 실시 예에 따른 전자 장치에서 딥러닝 모델에 마일스톤 식별자를 할당하는 동작을 설명하기 위한 순서도이다.
- [0079] 단계(500)에서는, 미리 설정된 QoS 데이터를 확인할 수 있다. QoS 데이터는 사용자에게 의해 요청되는 서비스 품질을 나타내는 지표로서, 딥러닝 모델을 통한 추론 과정이 수행되기 전에 미리 설정되어 저장될 수 있다.
- [0080] 단계(510)에서는, 확인된 QoS 데이터에 기초하여, 딥러닝 모델에 대한 압축을 수행할 수 있다.
- [0081] 단계(520)에서는, 압축이 수행된 딥러닝 모델에 대한 프로파일링을 수행하여, 딥러닝 모델 내 복수의 레이어들 각각에 대한 동작 시간을 측정할 수 있다.
- [0082] 단계(530)에서는, 측정 결과에 기초하여, 복수의 레이어들 중 적어도 하나의 레이어를 선택하고, 선택된 적어도 하나의 레이어에 마일스톤 식별자를 할당할 수 있다.
- [0083] 다양한 실시 예들에 따르면, 도 5에 개시된 동작들 중 일부는 생략되거나 복수 회 반복될 수 있다. 또한, 도 5에 개시된 동작들 각각은 일 실시 예로 보는 것이 타당하며, 어느 하나의 동작이 다른 하나의 동작에 종속되는 것으로 제한 해석될 수 없다.
- [0085] 도 6은 본 발명의 일 실시 예에 따른 전자 장치에서 딥러닝 모델을 통한 추론 과정을 수행할 때, 적어도 하나의 프로세서의 성능을 제어하는 방법을 설명하기 위한 순서도이다.
- [0086] 단계(600)에서는, 입력 데이터가 수신되면, 딥러닝 모델을 이용하여 수신된 입력 데이터에 대한 연산을 수행할 수 있다.
- [0087] 단계(610)에서는, 종료 식별자가 식별되는지 여부를 판단할 수 있다. 만약, 종료 식별자가 식별되면, 딥러닝 모델을 통한 연산은 종료될 수 있으며, 이 후 단계(660)이 수행될 수 있다. 반면에, 종료 식별자가 식별되지 않는다면, 단계(620)이 수행될 수 있다.
- [0088] 단계(620)에서는, 마일스톤 식별자가 식별되는지 여부를 판단할 수 있다. 만약, 마일스톤 식별자가 식별되면, 단계(630) 내지 단계(650)이 수행될 수 있다. 반면에, 마일스톤 식별자가 식별되지 않으면, 단계(600)에 따라 딥러닝 모델을 통한 연산이 속행될 수 있다.
- [0089] 단계(630)에서는, 입력 데이터에 대한 연산이 시작된 후 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장할 수 있다.
- [0090] 단계(640)에서는, 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득할 수 있다.
- [0091] 단계(650)에서는, 지연 값이 미리 설정된 값을 기준으로 임계 범위 내에 있지 않는 것에 응답하여, 적어도 하나의 프로세서의 주파수를 변경할 수 있다.
- [0092] 단계(660)에서는, 입력 데이터와 대응되는 출력 데이터로서, 종료 식별자가 식별된 시점까지 연산된 데이터를 제공할 수 있다.
- [0093] 다양한 실시 예들에 따르면, 도 6에 개시된 동작들 중 일부는 생략되거나 복수 회 반복될 수 있다. 또한, 도 6에 개시된 동작들 각각은 일 실시 예로 보는 것이 타당하며, 어느 하나의 동작이 다른 하나의 동작에 종속되는 것으로 제한 해석될 수 없다.
- [0095] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치는, 입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하고, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 레이어 관리부, 상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하고, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 성능 확인부, 상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하고, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 지연 제어부, 상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 리소스 관리부, 및 상기 주파수가 변경된 적어도 하나의

프로세서를 통해 상기 입력 데이터에 대한 연산이 수행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 수신된 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 입출력부를 포함할 수 있다.

- [0096] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치에 있어서, 상기 레이어 관리부는, 상기 복수의 레이어들 중 상기 입력 데이터에 대한 연산을 수행하는데 이용되는 제 1 복수의 레이어들을 선택하고, 상기 선택된 제 1 복수의 레이어들 각각의 타입을 식별하고, 상기 식별된 제 1 복수의 레이어들 각각의 타입과 대응되는 연산 공식을 이용하여 상기 입력 데이터에 대한 연산을 수행할 수 있다.
- [0097] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치는, 미리 설정된 QoS(quality of service) 데이터를 확인하고, 상기 확인된 QoS 데이터에 기초하여 상기 딥러닝 모델에 대한 압축을 수행하는 모델 압축부 및 상기 압축이 수행된 딥러닝 모델에 대한 프로파일링을 수행하여 상기 압축이 수행된 딥러닝 모델 내 상기 복수의 레이어들 각각에 대한 동작 시간을 측정하고, 상기 측정된 동작 시간을 이용하여 상기 복수의 레이어들 중 미리 설정된 시간 간격마다 동작되는 하나 이상의 레이어들을 선택하고, 상기 선택된 하나 이상의 레이어들 각각에 마일스톤 식별자를 할당하는 마일스톤 식별자 할당부를 더 포함할 수 있다.
- [0098] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치에 있어서, 상기 성능 확인부는, 상기 저장된 소요 시간의 상기 미리 설정된 목표 시간에 대한 비율을 상기 지연 값으로 획득할 수 있다.
- [0099] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치에 있어서, 상기 지연 제어부는, 상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 작으면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 하향 조정하도록 지시하는 제 1 제어 명령을 생성하고, 상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 크면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 상향 조정하도록 지시하는 제 2 제어 명령을 생성할 수 있다.
- [0100] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치에 있어서, 상기 지연 제어부는, 상기 획득된 지연 값이 상기 임계 범위를 벗어난 정도(degree)를 측정하고, 상기 측정된 정도에 기초하여 상기 적어도 하나의 프로세서의 성능이 하향 조정 또는 상향 조정되는 값을 결정하고, 상기 결정된 값에 따라 상기 제 1 제어 명령 또는 상기 제 2 제어 명령을 생성할 수 있다.
- [0101] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법은, 입력 데이터가 수신되면 딥러닝 모델을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 단계, 상기 입력 데이터에 대한 연산이 수행될 때 상기 딥러닝 모델에 포함된 복수의 레이어들 중 적어도 하나의 레이어에 미리 할당된 마일스톤 식별자를 식별하는 단계, 상기 마일스톤 식별자가 식별되는 것에 응답하여, 상기 입력 데이터에 대한 연산이 시작된 후 상기 마일스톤 식별자가 식별된 시점까지의 소요 시간을 측정하여 저장하는 단계, 상기 저장된 소요 시간 및 미리 설정된 목표 시간을 이용하여 지연 값을 획득하는 단계, 상기 획득된 지연 값이 미리 설정된 값을 기준으로 하여 임계 범위 내에 있는지 여부를 판단하는 단계, 상기 판단 결과에 기초하여 적어도 하나의 프로세서의 성능을 조정하기 위한 제어 명령을 생성하는 단계, 상기 생성된 제어 명령에 따라, 상기 전자 장치 내 적어도 하나의 프로세서의 주파수를 변경하는 단계, 및 상기 주파수가 변경된 적어도 하나의 프로세서를 통해 상기 입력 데이터에 대한 연산이 수행될 때 종료 식별자가 식별되는 것에 응답하여, 상기 입력 데이터와 대응되는 출력 데이터로서 상기 종료 식별자가 식별된 시점까지 연산된 데이터를 제공하는 단계를 포함할 수 있다.
- [0102] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법에 있어서, 상기 입력 데이터에 대한 연산을 수행하는 단계는, 상기 복수의 레이어들 중 상기 입력 데이터에 대한 연산을 수행하는데 이용되는 제 1 복수의 레이어들을 선택하는 단계, 상기 선택된 제 1 복수의 레이어들 각각의 타입을 식별하는 단계, 및 상기 식별된 제 1 복수의 레이어들 각각의 타입과 대응되는 연산 공식을 이용하여 상기 입력 데이터에 대한 연산을 수행하는 단계를 더 포함할 수 있다.
- [0103] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법은, 미리 설정된 QoS(quality of service) 데이터를 확인하고, 상기 확인된 QoS 데이터에 기초하여 상기 딥러닝 모델에 대한 압축을 수행하는 단계, 상기 압축이 수행된 딥러닝 모델에 대한 프로파일링을 수행하여 상기 압축이 수행된 딥러닝 모델 내 상기 복수의 레이어들 각각에 대한 동작 시간을 측정하는 단계, 상기 측정된 동작

시간을 이용하여 상기 복수의 레이어들 중 미리 설정된 시간 간격마다 동작되는 하나 이상의 레이어들을 선택하는 단계, 및 상기 선택된 하나 이상의 레이어들 각각에 마일스톤 식별자를 할당하는 단계를 더 포함할 수 있다.

[0104] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법에 있어서, 상기 지연 값을 획득하는 단계는, 상기 저장된 소요 시간의 상기 미리 설정된 목표 시간에 대한 비율을 상기 지연 값으로 획득하는 단계를 더 포함할 수 있다.

[0105] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법에 있어서, 상기 제어 명령을 생성하는 단계는, 상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 작으면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 하향 조정하도록 지시하는 제 1 제어 명령을 생성하는 단계, 및 상기 획득된 지연 값이 상기 미리 설정된 값인 1보다 크면서 상기 임계 범위를 벗어난 것으로 판단되는 것에 응답하여, 상기 전자 장치 내 상기 적어도 하나의 프로세서의 성능을 상향 조정하도록 지시하는 제 2 제어 명령을 생성하는 단계를 더 포함할 수 있다.

[0106] 본 문서에 개시된 다양한 실시 예들에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법에 있어서, 상기 제어 명령을 생성하는 단계는, 상기 획득된 지연 값이 상기 임계 범위를 벗어난 정도(degree)를 측정하는 단계, 상기 측정된 정도에 기초하여 상기 적어도 하나의 프로세서의 성능이 하향 조정 또는 상향 조정되는 값을 결정하는 단계, 및 상기 결정된 값에 따라 상기 제 1 제어 명령 또는 상기 제 2 제어 명령을 생성하는 단계를 더 포함할 수 있다.

[0107] 본 발명의 일 실시 예에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법은 컴퓨터와의 결합을 통해 실행시키기 위한 저장매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다.

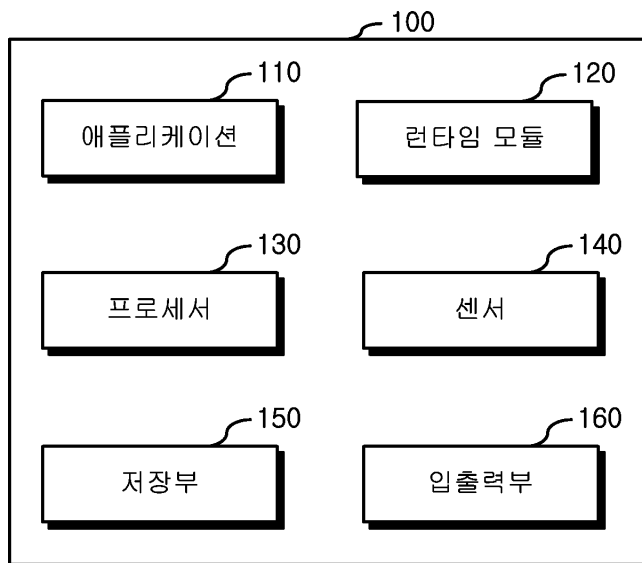
[0108] 또한, 본 발명의 일 실시 예에 따라 딥러닝 모델을 이용하여 연산을 수행하는 전자 장치를 제어하는 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0109] 이상과 같이 본 발명에서는 구체적인 구성 요소 등과 같은 특정 사항들과 한정된 실시 예 및 도면에 의해 설명되었으나 이는 본 발명의 보다 전반적인 이해를 돕기 위해서 제공된 것일 뿐, 본 발명은 상기의 실시 예에 한정되는 것은 아니며, 본 발명이 속하는 분야에서 통상적인 지식을 가진 자라면 이러한 기재로부터 다양한 수정 및 변형이 가능하다.

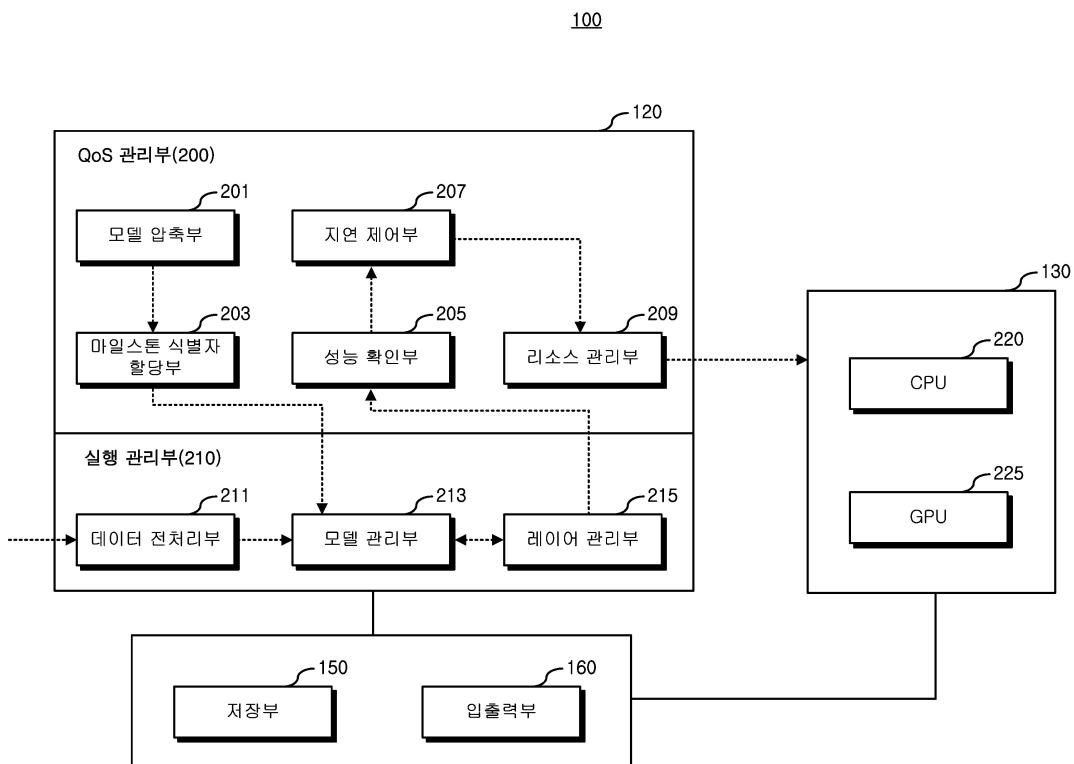
[0110] 따라서, 본 발명의 사상은 설명된 실시 예에 국한되어 정해져서는 아니 되며, 후술하는 특허청구범위뿐 아니라 이 특허청구범위와 균등하거나 등가적 변형이 있는 모든 것들은 본 발명 사상의 범주에 속한다고 할 것이다.

도면

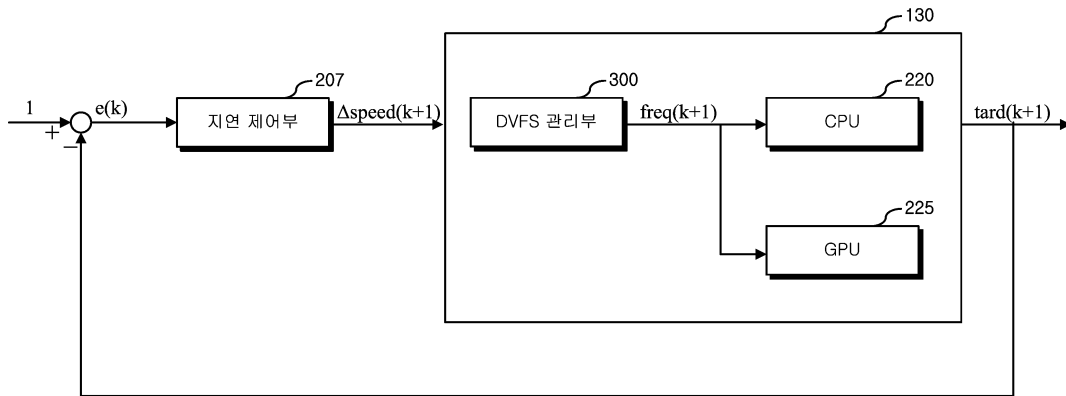
도면1



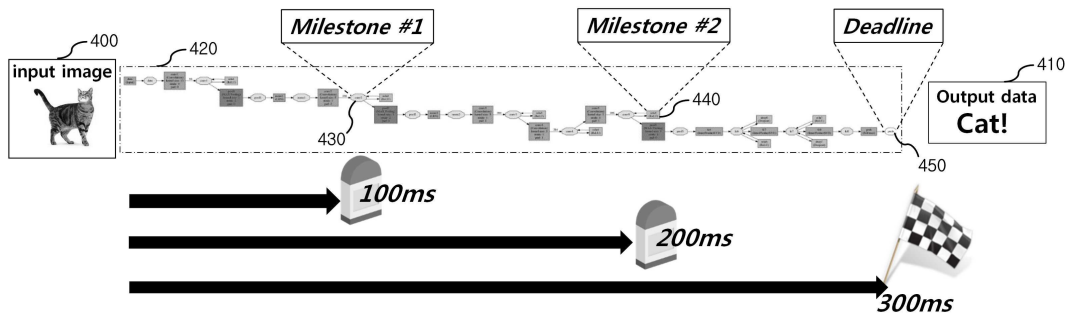
도면2



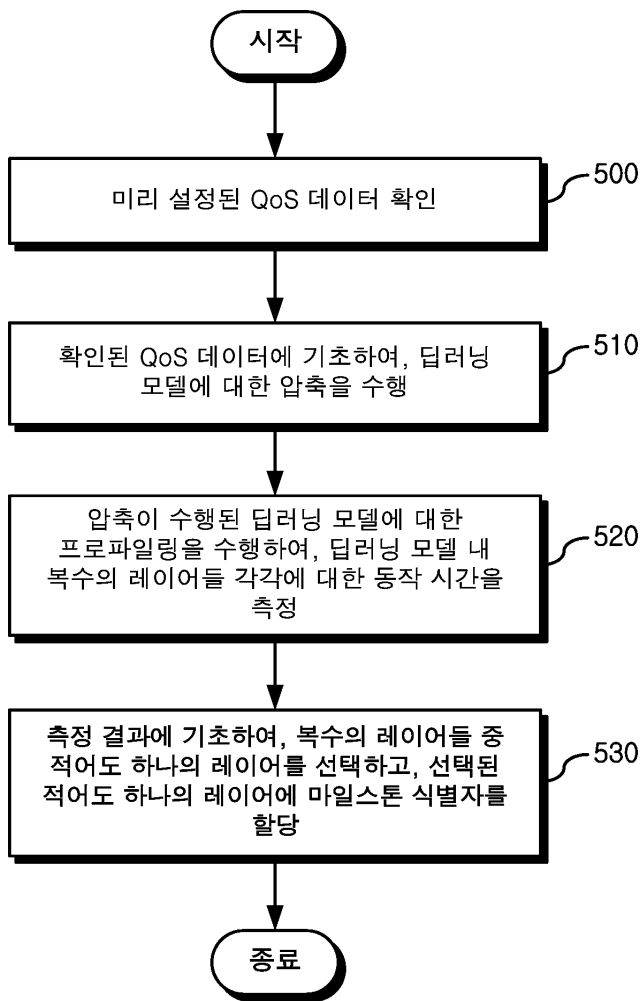
도면3



도면4



도면5



도면6

