(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0083716
(43) 공개일자 2023년06월12일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2023.01) *G06N 3/08* (2023.01)
(52) CPC특허분류
G06N 3/04 (2023.01)
G06N 3/082 (2023.01)
(21) 출원번호 10-2021-0171979
(22) 출원일자 2021년12월03일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
이원희
경기도 용인시 기흥구 금화로58번길 10, 403동 1405호 (상갈동, 금화마을주공그린빌)
(74) 대리인
특허법인 무한

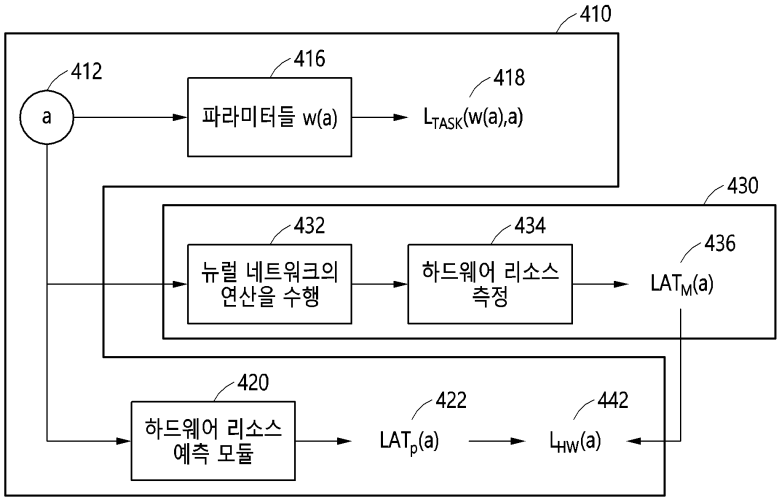
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **뉴럴 네트워크의 최적의 아키텍처를 탐색하는 장치 및 방법**

(57) 요 약

뉴럴 네트워크의 최적의 아키텍처를 탐색하는 방법 및 장치가 개시된다. 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 장치는 프로세서를 포함하고, 프로세서는, 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고, 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하고, 하드웨어 리소스 예측 모듈을 이용하여 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하고, 측정된 하드웨어 리소스와 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하고, 뉴럴 네트워크 손실과 하드웨어 리소스 손실에 기초하여 뉴럴 네트워크의 타겟 아키텍처를 결정할 수 있다.

대 표 도 - 도4



명 세 서

청구범위

청구항 1

뉴럴 네트워크의 최적의 아키텍처를 탐색하는 장치에 있어서,

프로세서를 포함하고,

상기 프로세서는,

상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고,

상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하고,

하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하고,

상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하고,

상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 뉴럴 네트워크의 타겟 아키텍처를 결정하는,

장치.

청구항 2

제1항에 있어서,

상기 하드웨어 리소스 예측 모듈은,

상기 후보 아키텍처의 파라미터들을 입력으로 하고, 상기 입력된 파라미터들을 기초로 상기 후보 아키텍처의 뉴럴 네트워크의 상기 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값을 출력하는 뉴럴 네트워크인,

장치.

청구항 3

제1항에 있어서,

상기 프로세서는,

상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스 간의 차이에 기초하여 상기 하드웨어 리소스 손실을 결정하고,

상기 하드웨어 리소스 손실이 최소가 되도록 상기 후보 아키텍처의 파라미터들을 업데이트하는,

장치.

청구항 4

제1항에 있어서,

상기 프로세서는,

상기 후보 아키텍처 및 상기 후보 아키텍처의 파라미터들에 따른 상기 뉴럴 네트워크 손실과 상기 후보 아키텍처의 파라미터들에 따른 상기 하드웨어 리소스 손실의 가중 합을 최적화 손실로 결정하고, 상기 최적화 손실을

최소로 만드는 최소로 만드는 상기 타겟 아키텍처를 결정하는,

장치.

청구항 5

제1항에 있어서,

상기 프로세서는,

상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실이 줄어들도록 하는 타겟 아키텍처와 타겟 파라미터들을 결정하는,

장치.

청구항 6

제1에 있어서,

상기 뉴럴 네트워크의 각 레이어별로 각 레이어가 가지는 후보 연산들 중 어느 하나의 후보 연산을 선택하는 것에 의해 상기 후보 아키텍처가 결정되는,

장치.

청구항 7

제1항에 있어서,

상기 하드웨어 리소스 예측 모듈에는 상기 뉴럴 네트워크의 각 레이어에서 수행 가능한 후보 연산들 중에서 상기 후보 아키텍처를 구성하는 선택된 후보 연산에 대한 정보가 입력되는,

장치.

청구항 8

제6항에 있어서,

상기 하드웨어 리소스는,

상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때의 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 중 적어도 하나를 포함하는,

장치.

청구항 9

제1항에 있어서,

상기 프로세서는,

상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실을 포함하는 최적화 손실을 결정하고,

상기 후보 아키텍처의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 상기 최적화 손실을 최소로 만드는 타겟 연산을 선택하는 것에 의해 상기 타겟 아키텍처를 결정하는,

장치.

청구항 10

제1항에 있어서,

상기 프로세서는,

상기 후보 아키텍처의 뉴럴 네트워크가 학습 데이터를 처리하여 도출한 결과 데이터와 검증 데이터(validation data) 간의 차이에 기초하여 상기 뉴럴 네트워크 손실을 결정하는,

장치.

청구항 11

증강 현실 제공 장치에 있어서,

타겟 아키텍처를 가지는 뉴럴 네트워크를 이용하여 처리 동작을 수행하는 프로세서를 포함하고,

상기 프로세서는,

상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고,

상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하고,

하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하고,

상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하고,

상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 타겟 아키텍처를 결정하는,

증강 현실 제공 장치.

청구항 12

제11항에 있어서,

상기 증강 현실 제공 장치는,

영상을 촬영하는 카메라를 더 포함하고,

상기 프로세서는,

상기 타겟 아키텍처를 가지는 뉴럴 네트워크를 이용하여 상기 영상을 처리하는 것에 의해 증강 현실 콘텐츠를 생성하는,

증강 현실 제공 장치.

청구항 13

제11항에 있어서,

상기 하드웨어 리소스 예측 모듈은,

상기 후보 아키텍처의 파라미터들을 입력으로 하고, 상기 입력된 파라미터들을 기초로 상기 후보 아키텍처의 뉴럴 네트워크의 상기 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값을 출력하는 뉴럴 네트워크인,

증강 현실 제공 장치.

청구항 14

제11항에 있어서,

상기 프로세서는,

상기 후보 아키텍처 및 상기 후보 아키텍처의 파라미터들에 따른 상기 뉴럴 네트워크 손실과 상기 후보 아키텍처의 파라미터들에 따른 상기 하드웨어 리소스 손실의 가중 합을 최소로 만드는 상기 타겟 아키텍처를 결정하는,

증강 현실 제공 장치.

청구항 15

뉴럴 네트워크의 최적의 아키텍처를 탐색하는 방법에 있어서,

상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하는 동작;

상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하는 동작;

하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하는 동작;

상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하는 동작; 및

상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 뉴럴 네트워크의 타겟 아키텍처를 결정하는 동작

을 포함하는 방법.

청구항 16

제15항에 있어서,

상기 하드웨어 리소스 예측 모듈은,

상기 후보 아키텍처의 파라미터들을 입력으로 하고, 상기 입력된 파라미터들을 기초로 상기 후보 아키텍처의 뉴럴 네트워크의 상기 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값을 출력하는 뉴럴 네트워크인,

방법.

청구항 17

제15항에 있어서,

상기 타겟 아키텍처를 결정하는 동작은,

상기 후보 아키텍처 및 상기 후보 아키텍처의 파라미터들에 따른 상기 뉴럴 네트워크 손실과 상기 후보 아키텍처의 파라미터들에 따른 상기 하드웨어 리소스 손실의 가중 합을 최소로 만드는 상기 타겟 아키텍처를 결정하는 동작

을 포함하는 방법.

청구항 18

제15항에 있어서,

상기 타겟 아키텍처를 결정하는 동작은,

상기 후보 아키텍처의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 타겟 연산을 선택하는 것에 의해 상기 타겟 아키텍처를 결정하는 동작

을 포함하는 방법.

청구항 19

제15항에 있어서,

상기 하드웨어 리소스는,

상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때의 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 중 적어도 하나를 포함하는,

방법.

청구항 20

제15항 내지 제19항 중 어느 한 항의 방법을 수행하기 위한 명령어를 포함하는 하나 이상의 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 기록 매체.

발명의 설명

기 술 분 야

[0001] 아래의 개시는 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 기술에 관한 것이다.

배 경 기 술

[0003] 뉴럴 아키텍처 탐색(neural architecture search, NAS)은 소정 목적의 뉴럴 네트워크의 최적의 아키텍처를 자동으로 찾기 위한 방법론 중 하나이다. NAS는 특정 문제를 해결하기 위한 가장 적합한 뉴럴 네트워크의 아키텍처의 구조 및 형태를 딥 러닝(deep learning)을 통해 탐색하는 방법이다. NAS에서의 뉴럴 네트워크는 탐색 공간(search space)라고 부르는, 미리 정의한 연산자 및 함수들로 구성된 프리미티브 연산들(primitive operations)을 선택 및 조합하여 생성될 수 있다. 여기서 연산자의 예시로는 컨볼루션(convolution), 풀링(pooling), 병합(concatenation), 스킵 연결(skip connection) 등이 있다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0005] 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 장치는, 프로세서를 포함하고, 상기 프로세서는, 상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고, 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하고, 하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하

고, 상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하고, 상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 뉴럴 네트워크의 타겟 아키텍처를 결정할 수 있다.

[0006] 상기 하드웨어 리소스 예측 모듈은, 상기 후보 아키텍처의 파라미터들을 입력으로 하고, 상기 입력된 파라미터들을 기초로 상기 후보 아키텍처의 뉴럴 네트워크의 상기 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값을 출력하는 뉴럴 네트워크일 수 있다.

[0007] 상기 프로세서는, 상기 후보 아키텍처 및 상기 후보 아키텍처의 파라미터들에 따른 상기 뉴럴 네트워크 손실과 상기 후보 아키텍처의 파라미터들에 따른 상기 하드웨어 리소스 손실의 가중 합을 최소로 만드는 상기 타겟 아키텍처를 결정할 수 있다.

[0008] 상기 하드웨어 리소스는, 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때의 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 중 적어도 하나를 포함할 수 있다.

[0009] 상기 프로세서는, 상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실을 포함하는 최적화 손실을 결정하고, 상기 후보 아키텍처의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 상기 최적화 손실을 최소로 만드는 타겟 연산을 선택하는 것에 의해 상기 타겟 아키텍처를 결정할 수 있다.

[0010] 일 실시예에 따른 증강 현실 제공 장치는, 타겟 아키텍처를 가지는 뉴럴 네트워크를 이용하여 처리 동작을 수행하는 프로세서를 포함하고, 상기 프로세서는, 상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고, 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하고, 하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하고, 상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하고, 상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 타겟 아키텍처를 결정할 수 있다.

[0011] 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 방법은, 상기 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하는 동작; 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정하는 동작; 하드웨어 리소스 예측 모듈을 이용하여 상기 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하는 동작; 상기 측정된 하드웨어 리소스와 상기 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정하는 동작; 및 상기 뉴럴 네트워크 손실과 상기 하드웨어 리소스 손실에 기초하여 상기 뉴럴 네트워크의 타겟 아키텍처를 결정하는 동작을 포함할 수 있다.

도면의 간단한 설명

[0013] 도 1은 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 탐색 프레임워크를 설명하기 위한 도면이다.

도 2는 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 탐색 장치의 구성을 도시하는 블록도이다.

도 3은 일 실시예에 따른 각 레이어의 후보 연산들 중 최적의 타겟 연산을 선택하는 과정을 설명하기 위한 도면이다.

도 4는 일 실시예에 따른 뉴럴 네트워크의 타겟 아키텍처를 결정하기 위한 탐색 과정을 설명하기 위한 도면이다.

도 5는 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 방법의 동작들을 설명하기 위한 흐름도이다.

도 6은 일 실시예에 따른 전자 장치의 구성을 도시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0014] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형

태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0015] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0016] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0017] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 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.

[0018] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0019] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

[0021] 도 1은 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 탐색 프레임워크를 설명하기 위한 도면이다.

[0022] 도 1을 참조하면,탐색 프레임워크(100)는 기계 학습(machine learning)을 통해 기본 뉴럴 네트워크(120)에 대한 최적의 아키텍처(architecture)(또는 뉴럴 네트워크 구조)를 탐색하는 프레임워크이다. 기본 뉴럴 네트워크(120)는 학습되기 이전의 뉴럴 네트워크(또는 학습되지 않은 뉴럴 네트워크)로서, 각 레이어들의 연산 및 파라미터(예: 연결 가중치) 등이 확정되지 않은 뉴럴 네트워크이다. 기본 뉴럴 네트워크(120)는 복수의 뉴럴 네트워크 레이어(또는 간단히 '레이어')들을 포함할 수 있다. 기본 뉴럴 네트워크(120)는 심층 뉴럴 네트워크(deep neural network, DNN), 컨볼루션 뉴럴 네트워크(convolutional neural network, CNN), 재귀적 뉴럴 네트워크(recurrent neural network, RNN), RBM(restricted boltzmann machine), DBN(deep belief network), BRDNN(bidirectional recurrent deep neural network), 심층 Q-네트워크(deep Q-networks) 또는 이들 중 둘 이상의 조합 중 하나일 수 있으나, 전술한 예에 한정되지 않는다. 기본 뉴럴 네트워크(120)는 하드웨어 구조 및/또는 소프트웨어 구조를 포함할 수 있다.

[0023] 탐색 프레임워크(100)는 데이터베이스(110)에 저장된 학습 데이터(training data)를 이용하여 기본 뉴럴 네트워크(120)에 대해 기계 학습을 수행할 수 있다. 기계 학습은 지도 학습(supervised learning) 또는 부분 지도 학습(partial supervised learning) 방식으로 수행될 수 있다.

[0024] 일 실시예에서, 탐색 프레임워크(100)는 지도 학습을 통해 기본 뉴럴 네트워크(120)를 학습시킬 수 있다. 탐색 프레임워크(100)는 확률적 경사 하강 기법(stochastic gradient descent)과 같은 조정 알고리즘 및 손실 함수를 이용하여 학습을 수행할 수 있다. 학습에 이용되는 학습 데이터는 뉴럴 네트워크에 입력되는 입력 데이터와 해당 입력 데이터에 대응하는 검증 데이터(validation data)를 포함할 수 있다. 기본 뉴럴 네트워크(120)는 학습 데이터에 포함된 입력 데이터를 처리하여 결과 데이터를 출력할 수 있다. 탐색 프레임워크(100)는 기본 뉴럴 네트워크(120)로부터 출력된 결과 데이터와 검증 데이터 간의 비교 결과에 기초하여 뉴럴 네트워크 손실(neural network loss)을 결정하고, 뉴럴 네트워크 손실을 최소화하는 최적의 아키텍처를 탐색할 수 있다.

[0025] 탐색 프레임워크(100)는 다중 목적(multiple objective)의 뉴럴 아키텍처 탐색(NAS) 기법을 수행하여 타겟 뉴럴 네트워크(130)를 위한 최적의 아키텍처를 탐색할 수 있다. 탐색 프레임워크(100)는 기본 뉴럴 네트워크(120)의 아키텍처를 샘플링하는 것이 아니라, 기본 뉴럴 네트워크(120)의 각 레이어(layer)마다 여러 후보 연산들(candidate operations)을 설정하고, 이들 후보 연산들 중에서 가장 적합한 후보 연산인 타겟 연산을 레이어마

다 선택하는 방식으로 최적의 아키텍처를 탐색할 수 있다. 이러한 탐색 방법을 통해 탐색 프레임워크(100)는 빠른 시간 내에 최적화를 수행할 수 있다.

[0026] 탐색 프레임워크(100)는 학습 과정을 통해 주어진 목적(예: 객체 분류, 객체 인식, 음성 인식 등)에 따른 최적의 아키텍처를 가지는 타겟 뉴럴 네트워크(130)를 도출할 수 있다. 최적의 아키텍처를 탐색하는 것은 뉴럴 네트워크의 각 레이어에서 수행되는 연산을 결정하고, 뉴럴 네트워크 파라미터들의 최적 값을 결정하는 것을 포함할 수 있다. 탐색 프레임워크(100)는 본 명세서에서 설명되는 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 장치(예: 도 2의 탐색 장치(200))에 의해 수행될 수 있다.

[0027] 탐색 프레임워크(100)는 타겟 뉴럴 네트워크(130)를 위한 최적의 아키텍처를 탐색하는데 있어 하드웨어 리소스 제한 사항(hardware resource constraint)을 고려할 수 있다. 탐색 프레임워크(100)는 뉴럴 네트워크가 수행하는 태스크(task)에 대한 검증 손실(validation loss) 뿐만 아니라 뉴럴 네트워크가 수행될 때 사용되는 하드웨어 리스소를 고려하여 최적화를 수행할 수 있다. 탐색 프레임워크(100)는 뉴럴 네트워크가 동작 시에 필요로 하는 하드웨어 리소스를 고려하여 타겟 뉴럴 네트워크(130)를 탐색할 수 있다. 하드웨어 리소스는 예를 들어 전력 소비, 메모리 요구량, 연산의 수(number of operations)(예: maultiply-accumulate(MAC) 연산의 수), 처리 시간, 및 GPU 점유율 등일 수 있다. 탐색 프레임워크(100)는 하나 또는 둘 이상의 하드웨어 리소스를 고려할 수 있고, 위 예로 든 하드웨어 리소스 이외에도 수치로 관측될 수 있는 하드웨어 리소스라면 어느 것이든 제한되지 않고 고려할 수 있다.

[0028] 탐색 프레임워크(100)는 기본 뉴럴 네트워크(120)에 대한 후보 아키텍처가 결정되면, 후보 아키텍처에 대한 뉴럴 네트워크 손실과 하드웨어 리소스에 대한 하드웨어 리소스 손실(hardware resource loss)을 결정하고, 뉴럴 네트워크 손실과 하드웨어 리소스 손실을 최소화하는 타겟 아키텍처를 탐색할 수 있다. 뉴럴 네트워크 손실과 하드웨어 리소스 손실은 타겟 아키텍처를 결정하기 위한 최적화 손실을 구성할 수 있다.

[0029] 탐색 프레임워크(100)는 하드웨어 리소스 손실을 결정할 때, 후보 아키텍처의 뉴럴 네트워크가 요구하는 하드웨어 리소스에 대한 실제 측정 값과 하드웨어 리소스 예측 모듈(예: 도 4의 하드웨어 리소스 예측 모듈(420))을 이용하여 도출된 하드웨어 리소스에 대한 예측 값에 기초하여 하드웨어 리소스 손실을 결정할 수 있다. 아래에서 보다 자세히 설명되지만, 하드웨어 리소스 예측 모듈은 현재 학습의 대상인 후보 아키텍처의 뉴럴 네트워크에 대한 하드웨어 리소스를 예측한 예측 값을 제공하는 모듈이다. 하드웨어 리소스 예측 모듈은 입력된 후보 아키텍처의 파라미터들에 기초하여 해당 후보 아키텍처의 뉴럴 네트워크가 필요로 하는 하드웨어 리소스에 대한 예측 값을 출력하도록 학습된 뉴럴 네트워크에 의해 구현될 수 있다. 하드웨어 리소스 예측 모듈은 미분 가능한(differentiable) 특성을 가지며, 하드웨어 리소스 예측 모듈을 통해 탐색 과정에서의 미분 가능성(differentiability)이 유지될 수 있다. 미분 가능성이 유지됨에 따라 종단간 학습(end-to-end learning)이 가능해질 수 있다. 탐색 프레임워크(100)는 하드웨어 리소스 예측 모듈을 통해 뉴럴 네트워크의 아키텍처에 대한 하드웨어 리소스를 최적화 손실에 반영할 수 있다.

[0030] 위와 같이, 탐색 프레임워크(100)는 하드웨어 리소스 제한 사항을 고려하여 뉴럴 네트워크의 최적의 아키텍처를 탐색할 수 있으며, 짧은 시간으로 최적화를 수행할 수 있다. 또한, 탐색 프레임워크(100)는 실제 측정한 하드웨어 리소스 측정 값을 고려하여 최적의 아키텍처를 탐색할 수 있다.

[0032] 도 2는 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 탐색 장치의 구성을 도시하는 블록도이다.

[0033] 도 2를 참조하면, 탐색 장치(200)는 뉴럴 네트워크에 대한 최적의 아키텍처를 탐색하는 장치로서, 도 1에서 설명한 탐색 프레임워크(100)를 수행할 수 있다. 탐색 장치(200)는 아키텍처 탐색과 관련하여 본 명세서에서 기술되거나 또는 도면에 도시된 하나 이상의 동작을 수행할 수 있다. 탐색 장치(200)는 프로세서(210) 및 메모리(220)를 포함할 수 있다. 저장 장치(230)는 아키텍처 탐색을 위한 데이터(예: 학습 데이터)를 저장하고, 학습에 이용되는 뉴럴 네트워크를 저장할 수 있다.

[0034] 메모리(220)는 탐색 장치(200)의 구성요소(예: 프로세서(210))에 의해 사용되는 다양한 데이터를 저장할 수 있다. 데이터는, 예를 들어, 소프트웨어 및, 이와 관련된 명령에 대한 입력 데이터 또는 출력 데이터를 포함할 수 있다. 메모리(220)는 휘발성 메모리 및 비휘발성 메모리 중 하나 이상을 포함할 수 있다.

[0035] 프로세서(210)는 탐색 장치(200)의 전체적인 동작을 제어하며, 탐색 장치(200)의 동작을 수행하기 위한 인스트럭션들을 실행한다. 프로세서(210)는 예를 들어 소프트웨어를 실행하여 프로세서(210)에 연결된 탐색 장치

(200)의 적어도 하나의 다른 구성요소(예: 하드웨어 또는 소프트웨어 구성요소)를 제어할 수 있고, 다양한 데이터 처리 또는 연산을 수행할 수 있다.

[0036] 일 실시예에 따르면, 데이터 처리 또는 연산의 적어도 일부로서, 프로세서(210)는 명령 또는 데이터를 메모리(220)에 저장하고, 메모리(220)에 저장된 명령 또는 데이터를 처리하고, 결과 데이터를 메모리(220)에 저장할 수 있다. 프로세서(210)는 메인 프로세서(예: 중앙 처리 장치 또는 어플리케이션 프로세서) 또는 이와는 독립적으로 또는 함께 운영 가능한 보조 프로세서(예: 그래픽 처리 장치, 신경망 처리 장치(NPU: neural processing unit), 이미지 시그널 프로세서, 센서 허브 프로세서, 또는 커뮤니케이션 프로세서)를 포함할 수 있다.

[0037] 프로세서(210)는 학습 데이터를 이용하여 뉴럴 네트워크(예: 도 1의 기본 뉴럴 네트워크(120))의 후보 아키텍처에 대한 학습을 진행하고, 뉴럴 네트워크 손실을 결정할 수 있다. 학습되기 이전의 뉴럴 네트워크는 각각 하나 이상의 인공 뉴런을 포함하는 복수의 레이어들을 포함하고, 각 레이어는 수행 가능한 후보 연산들이 미리 정의되어 있을 수 있다. 후보 연산들은 예를 들어 3X3 커널(kernel) 기반의 컨볼루션 연산, 5X5 커널 기반의 컨볼루션 연산, 풀링(pooling) 연산 등을 포함할 수 있으나, 이에 제한되는 것은 아니다. 뉴럴 네트워크의 각 레이어별로 각 레이어가 가지는 후보 연산들 중 어느 하나의 후보 연산을 선택하는 것에 의해 후보 아키텍처가 결정될 수 있다. 프로세서(210)는 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정할 수 있다. 프로세서(210)는 후보 아키텍처의 뉴럴 네트워크가 학습 데이터를 처리하여 도출한 결과 데이터와 검증 데이터 간의 차이에 기초하여 뉴럴 네트워크 손실을 결정할 수 있다. 뉴럴 네트워크 손실은 미리 정의된 손실 함수에 의해 결정될 수 있다.

[0038] 프로세서(210)는 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구(또는 사용)되는 하드웨어 리소스를 측정할 수 있다. 측정되는 하드웨어 리소스는 예를 들어 후보 아키텍처의 뉴럴 네트워크가 동작할 때의 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 중 하나 또는 둘 이상을 포함할 수 있으나, 이에 제한되는 것은 아니다. 프로세서(210)는 하드웨어 리소스를 측정하여 하드웨어 리소스 측정 값을 결정할 수 있다.

[0039] 프로세서(210)는 하드웨어 리소스 예측 모듈(예: 도 4의 하드웨어 리소스 예측 모듈(420))을 이용하여 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측할 수 있다. 하드웨어 리소스 예측 모듈은 후보 아키텍처의 파라미터들(예: 각 레이어의 후보 연산들 중에서 선택된 후보 연산에 대한 정보)을 입력으로 하고, 입력된 파라미터들을 기초로 후보 아키텍처의 뉴럴 네트워크의 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값을 출력하는 뉴럴 네트워크일 수 있다.

[0040] 프로세서(210)는 측정된 하드웨어 리소스와 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정할 수 있다. 프로세서(210)는 측정된 하드웨어 리소스와 예측된 하드웨어 리소스 간의 차이와 미리 정의된 손실 함수에 기초하여 하드웨어 리소스 손실을 결정할 수 있다. 예를 들어, 프로세서(210)는 후보 아키텍처의 뉴럴 네트워크가 동작할 때 실제로 측정된 처리 시간 측정 값과 하드웨어 리소스 예측 모델로부터 출력된 처리 시간 예상 값 간의 차이를 손실 함수에 적용하여 하드웨어 리소스 손실을 결정할 수 있다.

[0041] 프로세서(210)는 뉴럴 네트워크 손실과 하드웨어 리소스 손실에 기초하여 뉴럴 네트워크의 타겟 아키텍처를 결정할 수 있다. 프로세서(210)는 뉴럴 네트워크 손실과 하드웨어 리소스 손실이 줄어들도록 하는 타겟 아키텍처와 타겟 파라미터들을 결정할 수 있다. 프로세서(210)는 하드웨어 리소스 손실이 최소가 되도록 후보 아키텍처의 파라미터들을 업데이트할 수 있다. 프로세서(210)는 뉴럴 네트워크 손실과 하드웨어 리소스 손실을 포함하는 최적화 손실을 결정하고, 후보 아키텍처의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 최적화 손실을 최소로 만드는 타겟 연산을 선택하는 것에 의해 타겟 아키텍처를 결정할 수 있다. 프로세서(210)는 뉴럴 네트워크 손실과 하드웨어 리소스 손실을 가장 최소화하는 각 레이어의 타겟 연산을 선택하고, 뉴럴 네트워크의 파라미터들을 업데이트할 수 있다. 프로세서(210)는 후보 아키텍처 및 후보 아키텍처의 파라미터들에 따른 뉴럴 네트워크 손실과 후보 아키텍처의 파라미터들에 따른 하드웨어 리소스 손실의 가중 합을 최적화 손실로 결정하고, 해당 최적화 손실을 최소로 만드는 타겟 아키텍처를 결정할 수 있다.

[0042] 이상에서 설명된 탐색 장치(200)에 의해 수행되는 동작들은 웨어러블 장치, 스마트폰과 같은 모바일 장치 뿐만 아니라 임베디드 시스템(embedded system)에서 작동할 수 있는 뉴럴 네트워크 기반의 알고리즘에 다양하게 응용될 수 있다.

[0044] 도 3은 일 실시예에 따른 각 레이어의 후보 연산들 중 최적의 타겟 연산을 선택하는 과정을 설명하기 위한 도면이다.

[0045] 도 3을 참조하면, 단계(310)에서 학습되기 이전의 뉴럴 네트워크(예: 도 1의 기본 뉴럴 네트워크(120))는 복수의 레이어들(312, 314, 316, 318)을 포함하고, 각 레이어들(312, 314, 316, 318)은 수행 가능한 후보 연산들이 정의되어 있을 수 있다. 도시된 실시예에서, 뉴럴 네트워크(310)는 각 레이어들(312, 314, 316, 318)에 대해 3개의 후보 연산들을 가질 수 있는 것으로 정의되어 있다. 레이어들 간에 수행되는 후보 연산들은 서로 다른 연산 방식일 수 있다. 예를 들어, 레이어(312)와 레이어(314) 사이에서 수행되는 후보 연산들은 서로 다른 방식의 연산들일 수 있다.

[0046] 탐색 장치(예: 도 2의 탐색 장치(200))는 학습 과정에서 이들 후보 연산들 중 최적의 후보 연산인 타겟 연산을 선택할 수 있다. 단계(320)에서, 탐색 장치는 각 레이어들(312, 314, 316, 318)별로 후보 연산들 중 어느 하나의 후보 연산(322, 324, 326, 328, 329)을 선택하고, 선택한 후보 연산들(322, 324, 326, 328, 329)의 조합으로 구성된 후보 아키텍처에 대한 최적화 손실을 결정할 수 있다. 탐색 장치는 탐색 공간에서 각 레이어들(312, 314, 316, 318)의 후보 연산들을 여러 번 조합하고, 각각의 조합에서의 최적화 손실을 계산하여 최적화 손실을 최소로 하는 후보 연산(다른 말로, 타겟 연산)들의 조합을 결정할 수 있다. 다양한 조합에 대한 학습 과정이 완료되면, 각 레이어들(312, 314, 316, 318)별로 선택된 타겟 연산을 기초로 타겟 아키텍처를 결정할 수 있다. 타겟 아키텍처는 각 레이어의 타겟 연산을 포함한다.

[0048] 도 4는 일 실시예에 따른 뉴럴 네트워크의 타겟 아키텍처를 결정하기 위한 탐색 과정을 설명하기 위한 도면이다.

[0049] 도 4를 참조하면, 뉴럴 네트워크의 타겟 아키텍처의 탐색 과정에서는 뉴럴 네트워크의 후보 아키텍처 a(412)가 주어지면, 후보 아키텍처 a(412)의 파라미터들 w(a)(416)이 결정될 수 있다. 후보 아키텍처 a(412)는 각 레이어들별로 선택된 후보 연산의 집합을 포함하는 뉴럴 네트워크 구조를 나타낸다.

[0050] 후보 아키텍처 a(412)의 파라미터들은 뉴럴 네트워크의 각 레이어들의 후보 연산들 중에서 선택된 후보 연산에 대한 파라미터와 선택된 각 후보 연산에 대한 연산 특성을 나타내는 파라미터를 포함할 수 있다. 예를 들어, 뉴럴 네트워크에 포함된 어느 레이어의 후보 연산들이 3X3 커널 기반의 컨볼루션 연산 및 5X5 커널 기반의 컨볼루션 연산이 있다고 가정하면, 후보 아키텍처 a(412)의 파라미터들은 두 컨볼루션 연산들 중 어느 컨볼루션 연산이 선택되는지를 나타내는 파라미터와 선택된 컨볼루션 연산의 커널 파라미터(kernel parameter) 등을 포함할 수 있다. 여기서, 컨볼루션 연산은 컨볼루션 레이어로서 구현될 수도 있다.

[0051] 탐색 장치(예: 도 2의 탐색 장치(200))는 파라미터들 w(a)(416)에 기초하여 후보 아키텍처 a(412)의 뉴럴 네트워크의 특정 태스크(task)에 대한 뉴럴 네트워크 손실 $L_{TASK}(w(a), a)(418)$을 결정할 수 있다. 뉴럴 네트워크 손실 $L_{TASK}(w(a), a)(418)$은 뉴럴 네트워크가 수행하는 태스크의 손실을 최소화하기 위한 손실이다.

[0052] 탐색 장치는 후보 아키텍처 a(412)를 가지는 뉴럴 네트워크의 연산을 수행(432)할 수 있고, 해당 연산의 수행 과정에서 요구되는 전체 뉴럴 네트워크의 하드웨어 리소스를 실제로 측정(434)할 수 있다. 측정되는 하드웨어 리소스는 예를 들어 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 등을 포함할 수 있다. 하드웨어 리소스의 측정(434)을 통해 후보 아키텍처 a(412)에 대한 하드웨어 리소스 측정 값 $LAT_M(a)(436)$이 결정될 수 있다. 실시예에 따라, 맥스 연산(max operation)을 통해 결정된 아키텍처에 대한 뉴럴 네트워크 연산을 수행하는 것을 통해 하드웨어 리소스가 측정될 수도 있다. 이러한 과정들을 포함하는 과정(430)은 미분 가능성이 성립하지 않기 때문에 해당 과정(430)에 대해서는 포워드(forward) 및 백워드(backward)를 정의할 수 없다. 이를 해결하기 위해 하드웨어 리소스 예측 모듈(420)이 이용될 수 있다.

[0053] 탐색 장치는 하드웨어 리소스 예측 모듈(420)을 이용하여 후보 아키텍처 a(412)를 가지는 뉴럴 네트워크이 연산을 수행할 때 요구할 것으로 예상되는 전체 뉴럴 네트워크의 하드웨어 리소스를 예측할 수 있다. 하드웨어 리소스 예측 모듈(420)을 통해 후보 아키텍처 a(412)에 대한 하드웨어 리소스 예측 값 $LAT_P(a)(422)$이 결정될 수 있다. 하드웨어 리소스 예측 모듈(420)은 후보 아키텍처 a(412)의 파라미터들을 입력으로 하고, 입력된 파라미터들을 기초로 후보 아키텍처 a(412)의 뉴럴 네트워크의 하드웨어 리소스를 예측하여 하드웨어 리소스 예측 값 $LAT_P(a)(422)$을 출력할 수 있다. 하드웨어 리소스 예측 모듈(420)에서의 연산은 미분 가능한 연산으로 이루어질 수 있다.

[0054] 하드웨어 리소스 예측 모듈(420)은 뉴럴 네트워크의 아키텍처의 파라미터들을 기초로 해당 아키텍처가 요구하거나 사용할 것으로 예측되는 하드웨어 리소스의 예측 값을 출력하도록 학습 과정을 통해 학습된 뉴럴 네트워크일

수 있다. 다만, 하드웨어 리소스 예측 모듈(420)은 뉴럴 네트워크 이외에, 후보 아키텍처 a(412)에 기초하여 뉴럴 네트워크의 하드웨어 리소스를 예측할 수 있는 다른 수단에 의해서도 구현될 수 있다.

[0055] 탐색 장치는 하드웨어 리소스 측정 값 $LAT_M(a)$(436)과 하드웨어 리소스 예측 값 $LAT_P(a)$(422)에 기초하여 후보 아키텍처 a(412)에 대한 리소스 손실 $L_{HW}(a)$(442)을 결정할 수 있다. 하드웨어 리소스 측정 값 $LAT_M(a)$(436)과 하드웨어 리소스 예측 값 $LAT_P(a)$(422) 간의 차이가 최소화되도록 리소스 손실 $L_{HW}(a)$(442)이 정의될 수 있다.

[0056] 일 실시예에서, 리소스 손실 $L_{HW}(a)$(442)은 하드웨어 리소스 측정 값 $LAT_M(a)$(436)과 하드웨어 리소스 예측 값 $LAT_P(a)$(422) 간의 차이에 의한 손실을 나타내는 $L_{HW1}(a)$과 하드웨어 리소스의 최적화를 위한 요소(예: 레이턴시(latency)를 최소화하기 위한 요소)인 $L_{HW2}(a)$에 기초하여 결정될 수 있다. $L_{HW1}(a)$와 $L_{HW2}(a)$는 각각 다음의 수학식 1 및 수학식 2에 의해 결정될 수 있다.

수학식 1

[0058] $$L_{HW1}(a) = (LAT_M(a) - LAT_P(a))^2$$

수학식 2

[0060] $$L_{HW2}(a) = (LAT_M(a))^2$$

[0062] 리소스 손실 $L_{HW}(a)$(442)은 예를 들어 다음의 수학식 3과 같이 $L_{HW1}(a)$와 $L_{HW2}(a)$ 간의 가중합으로서 결정될 수 있다.

수학식 3

[0063] $$L_{HW}(a) = L_{HW1}(a) + w \times L_{HW2}(a)$$

[0065] 여기서, w는 $L_{HW2}(a)$에 적용되는 가중치로서 예를 들어 미리 설정된 상수일 수 있다. 실시예에 따라, $L_{HW1}(a)$에만 가중치가 적용될 수도 있고, $L_{HW1}(a)$와 $L_{HW2}(a)$에 각각 서로 다른 가중치가 적용될 수도 있다.

[0066] 탐색 장치는 하드웨어 리소스 측정 값 $LAT_M(a)$(436)과 하드웨어 리소스 예측 값 $LAT_P(a)$(422) 간의 차이를 미리 정의된 손실 함수에 적용하여 하드웨어 리소스 손실 $L_{HW}(a)$(442)을 결정할 수 있다.

[0067] 탐색 장치는 뉴럴 네트워크 손실 $L_{TASK}(w(a), a)$(418)과 하드웨어 리소스 손실 $L_{HW}(a)$(442)을 포함하는 최적화 손실을 결정하고, 후보 아키텍처 a(412)의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 최적화 손실을 최소로 만드는 타겟 연산을 선택하는 것에 의해 타겟 아키텍처를 결정할 수 있다. 예를 들어, 타겟 아키텍처는 다음의 수학식 4과 같이 최적화 손실 $L_{TASK}(w(a), a) + \lambda \cdot L_{HW}(a)$ 을 최소로 만드는 후보 아키텍처 a와 후보 아키텍처 a의 파라미터들 w(a)을 탐색하는 것에 의해 결정될 수 있다. 수학식 4의 예에서

최적화 손실은 하드웨어 리소스 손실 $L_{HW}(a)$(442)에 가중치 λ 가 적용된 뉴럴 네트워크 손실 $L_{TASK}(w(a), a)$(418)과 하드웨어 리소스 손실 $L_{HW}(a)$(442) 간의 가중합(weighted sum)으로서 결정될 수 있다.

수학식 4

[0068]
$$\min_{a}\min_{w} L_{TASK}(w(a), a) + \lambda \cdot L_{\mathrm{HW}}(a)$$

[0069] 탐색 장치는 적은 탐색 시간으로 뉴럴 네트워크의 타겟 아키텍처를 탐색할 수 있고, 타겟 아키텍처를 선정하는 데 있어 뉴럴 네트워크가 필요로 하는 하드웨어 리소스를 최적화 제한 사항으로서 고려할 수 있다. 도 4에서 과정(410)은 미분 가능한 특성을 가지고, 과정(430)은 미분 불가능한(non-differentiable) 특성을 가진다. 탐색 장치는 뉴럴 네트워크의 각 레이어들의 연산에 기초한 전체 뉴럴 네트워크의 하드웨어 리소스를 뉴럴 네트워크로 구현될 수 있는 하드웨어 리소스 예측 모듈(420)을 이용하여 예측함으로써 미분 가능성을 유지할 수 있다. 미분 가능성을 유지함으로써 종단간 학습이 가능해 진다. 또한, 상술된 탐색 과정은 전체 뉴럴 네트워크의 하드웨어 리소스 제한 사항을 반영하여 뉴럴 네트워크의 아키텍처를 미분 가능하게 최적화할 수 있고, 한번의 학습 과정을 통해 타겟 아키텍처를 찾을 수 있으므로 최적화 시간이 짧은 장점을 가진다. 그리고, 타겟 아키텍처의 탐색 과정에서 일어나는 포워드 및 백워드 간의 일관성(consistency)도 유지될 수 있다.

[0071] 도 5는 일 실시예에 따른 뉴럴 네트워크의 최적의 아키텍처를 탐색하는 방법의 동작들을 설명하기 위한 흐름도이다. 해당 방법의 동작들은 도 2의 탐색 장치(200)에 의해 수행될 수 있다.

[0072] 도 5를 참조하면, 동작(510)에서 탐색 장치는 뉴럴 네트워크(예: 도 1의 기본 뉴럴 네트워크(120))의 후보 아키텍처를 선택할 수 있다. 탐색 장치는 뉴럴 네트워크의 각 레이어에 대해 정의된 후보 연산들 중에서 어느 하나의 후보 연산을 선택하는 것에 의해 후보 아키텍처를 선택할 수 있다.

[0073] 동작(520)에서, 탐색 장치는 뉴럴 네트워크의 후보 아키텍처에 대한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정할 수 있다. 탐색 장치는 학습 데이터를 이용하여 뉴럴 네트워크의 후보 아키텍처에 대한 학습을 진행하고, 뉴럴 네트워크 손실을 결정할 수 있다. 탐색 장치는 후보 아키텍처의 뉴럴 네트워크가 학습 데이터를 처리하여 도출한 결과 데이터와 검증 데이터 간의 차이에 기초하여 뉴럴 네트워크 손실을 결정할 수 있다. 후보 아키텍처의 뉴럴 네트워크로부터 도출된 결과 데이터와 목적하는 검증 데이터 간의 차이가 커질수록 뉴럴 네트워크 손실은 커질 수 있다.

[0074] 동작(530)에서, 탐색 장치는 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정할 수 있다. 측정되는 하드웨어 리소스는 예를 들어 후보 아키텍처의 뉴럴 네트워크가 동작할 때의 전력 소비, 메모리 요구량, 연산의 수 및 처리 시간 중 하나 또는 둘 이상을 포함할 수 있으나, 이에 제한되는 것은 아니다.

[0075] 동작(540)에서, 탐색 장치는 하드웨어 리소스 예측 모듈(예: 도 4의 하드웨어 리소스 예측 모듈(420))을 이용하여 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측할 수 있다. 하드웨어 리소스 예측 모델에는 뉴럴 네트워크의 각 레이어에서 수행 가능한 후보 연산들 중에서 후보 아키텍처를 구성하는 선택된 후보 연산에 대한 정보가 입력될 수 있고, 하드웨어 리소스 예측 모델은 입력된 정보에 기반하여 해당 뉴럴 네트워크가 필요로 할 하드웨어 리소스의 예측 값을 제공할 수 있다.

[0076] 동작(550)에서, 탐색 장치는 측정된 하드웨어 리소스와 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정할 수 있다. 탐색 장치는 측정된 하드웨어 리소스와 예측된 하드웨어 리소스 간의 차이와 미리 정의된 손실 함수에 기초하여 하드웨어 리소스 손실을 결정할 수 있다

[0077] 동작(560)에서, 탐색 장치는 뉴럴 네트워크 손실과 하드웨어 리소스 손실에 기초하여 뉴럴 네트워크의 타겟 아키텍처와 타겟 파라미터들을 결정할 수 있다. 탐색 장치는 뉴럴 네트워크 손실과 하드웨어 리소스 손실을 포함하는 최적화 손실을 최소로 만드는 타겟 아키텍처와 타겟 파라미터들을 결정할 수 있다. 탐색 장치는 후보 아키텍처의 뉴럴 네트워크에 포함된 각 레이어들의 후보 연산들 중에서 최적화 손실을 최소로 만드는 타겟 연산을 선택하는 것에 의해 타겟 아키텍처를 결정할 수 있다. 탐색 장치는 후보 아키텍처 및 후보 아키텍처의 파라미

터들에 따른 뉴럴 네트워크 손실과 후보 아키텍처의 파라미터들에 따른 하드웨어 리소스 손실의 가중 합을 최적화 손실로 결정하고, 해당 최적화 손실을 최소로 만드는 타겟 아키텍처를 결정할 수 있다.

[0079] 도 6은 일 실시예에 따른 전자 장치의 구성을 도시하는 도면이다.

[0080] 도 6을 참조하면, 전자 장치(600)는 다양한 형태의 전자 장치일 수 있다. 예를 들어, 전자 장치(600)는 웨어러블 장치(예: AR 글래스와 같은 증강 현실 제공 장치, HMD(head mounted display)) 스마트폰, 태블릿 컴퓨터, 넷북, 랩탑, 제품 검사 장치, 퍼스널 컴퓨터 또는 서버일 수 있으나, 이에 제한되지 않는다.

[0081] 전자 장치(600)는 프로세서(610), 저장 장치(620), 카메라(630), 센서(640), 출력 장치(650) 및 통신 장치(660)를 포함할 수 있다. 전자 장치(600)의 각 컴포넌트들 중 적어도 일부는 주변 기기들 간의 통신 인터페이스(670)(예: 버스(bus), GPIO(general purpose input and output), SPI(serial peripheral interface), 또는 MIPI(mobile industry processor interface))을 통해 서로 연결되고 신호(예: 명령 또는 데이터)를 상호간에 교환할 수 있다.

[0082] 프로세서(610)는 전자 장치(600)의 전체적인 동작을 제어하며, 전자 장치(600) 내에서 실행하기 위한 기능 및 인스트럭션을 실행한다. 프로세서(610)는 도 1 내지 도 5를 통하여 전술한 탐색 장치(예: 도 2의 탐색 장치(200))의 동작들을 수행할 수 있다.

[0083] 메모리(620)는 프로세서(610)에 의해 실행 가능한 인스트럭션들과 입/출력되는 데이터를 저장할 수 있다. 메모리(620)는 RAM, DRAM, SRAM과 같은 휘발성 메모리 및/또는 ROM, 플래쉬 메모리와 같은 이 기술 분야에서 알려진 비휘발성 메모리를 포함할 수 있다.

[0084] 카메라(630)는 영상을 촬영할 수 있다. 카메라(630)는 예를 들어 컬러 영상, 흑백 영상, 그레이 영상, 적외선 영상 또는 깊이 영상 등을 획득할 수 있다.

[0085] 센서(640)는 전자 장치(600)의 작동 상태(예: 전력 또는 온도), 또는 외부의 환경 상태(예: 사용자 상태)를 감지하고, 감지된 상태에 대응하는 전기 신호 또는 데이터 값을 생성할 수 있다. 센서(640)는 예를 들어 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, 또는 조도 센서를 포함할 수 있다.

[0086] 출력 장치(650)는 시각적, 청각적 또는 촉각적인 채널을 통해 사용자에게 전자 장치(600)의 출력을 제공할 수 있다. 출력 장치(660)는 예를 들어 액정 디스플레이나 LED/OLED 디스플레이, 마이크로 엘이디(micro light emitting diode, micro LED), 터치 스크린, 스피커, 진동 발생 장치 또는 사용자에게 출력을 제공할 수 있는 임의의 다른 장치를 포함할 수 있다.

[0087] 통신 장치(660)는 전자 장치(600)와 외부의 장치 간의 직접(예: 유선) 통신 채널 또는 무선 통신 채널의 수립, 및 수립된 통신 채널을 통한 통신 수행을 지원할 수 있다. 일 실시예에 따르면, 통신 모듈(600)은 무선 통신 모듈 (예: 셀룰러 통신 모듈, 근거리 무선 통신 모듈, 또는 GNSS(global navigation satellite system) 통신 모듈) 또는 유선 통신 모듈 (예: LAN(local area network) 통신 모듈, 또는 전력선 통신 모듈)을 포함할 수 있다. 무선 통신 모듈은 근거리 통신 네트워크(예: 블루투스, WiFi(wireless fidelity) direct 또는 IrDA(infrared data association)) 또는 원거리 통신 네트워크(예: 레거시 셀룰러 네트워크, 5G 네트워크, 차세대 통신 네트워크, 인터넷, 또는 컴퓨터 네트워크(예: LAN 또는 WAN))를 통하여 외부의 장치와 통신할 수 있다.

[0088] 일 실시예에서, 전자 장치(600)는 뉴럴 네트워크 기반의 알고리즘을 이용하는 증강 현실 제공 장치(예: AR(augmented reality) 글래스)일 수 있다. 증강 현실 제공 장치는 사용자의 안면에 착용되어 사용자에게 증강 현실 서비스 및/또는 가상 현실 서비스와 관련된 콘텐츠를 제공할 수 있다. 프로세서(610)는 타겟 아키텍처를 가지는 뉴럴 네트워크를 이용하여 처리 동작을 수행할 수 있다. 카메라(630)는 증강 현실 콘텐츠의 생성을 위한 영상을 촬영할 수 있고, 프로세서(610)는 타겟 아키텍처를 가지는 뉴럴 네트워크를 이용하여 영상을 처리하는 것에 의해 증강 현실 콘텐츠를 생성할 수 있다. 예를 들어, 프로세서(610)는 카메라(630)를 통해 획득된 영상에서 특정한 객체를 인식하고, 인식한 객체 영역 또는 객체 주변 영역에 가상의 콘텐츠를 중첩하여 표현하는 것에 의해 증강 현실 콘텐츠를 생성할 수 있다.

[0089] 프로세서(610)는 도 2 및 도 5에서 설명한 것과 같은 과정을 통해 뉴럴 네트워크의 타겟 아키텍처를 결정할 수 있다. 예를 들어, 프로세서(610)는 뉴럴 네트워크(예: 도 1의 기본 뉴럴 네트워크(120))의 후보 아키텍처에 대

한 파라미터들에 기초하여 뉴럴 네트워크 손실을 결정하고, 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구되는 하드웨어 리소스를 측정할 수 있다. 프로세서(610)는 하드웨어 리소스 예측 모듈(예: 도 4의 하드웨어 리소스 예측 모듈(420))을 이용하여 후보 아키텍처의 뉴럴 네트워크가 동작할 때에 요구될 하드웨어 리소스를 예측하고, 측정된 하드웨어 리소스와 예측된 하드웨어 리소스에 기초하여 하드웨어 리소스 손실을 결정할 수 있다. 프로세서(610)는 뉴럴 네트워크 손실과 하드웨어 리소스 손실에 기초하여 타겟 아키텍처를 결정할 수 있다. 프로세서(610)는 후보 아키텍처 및 후보 아키텍처의 파라미터들에 따른 뉴럴 네트워크 손실과 후보 아키텍처의 파라미터들에 따른 하드웨어 리소스 손실의 가중 합을 기초로 최적화 손실을 결정하고, 최적화 손실을 최소로 만드는 타겟 아키텍처와 후보 아키텍처의 파라미터들을 결정할 수 있다.

[0091] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0092] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0093] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 저장할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 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.

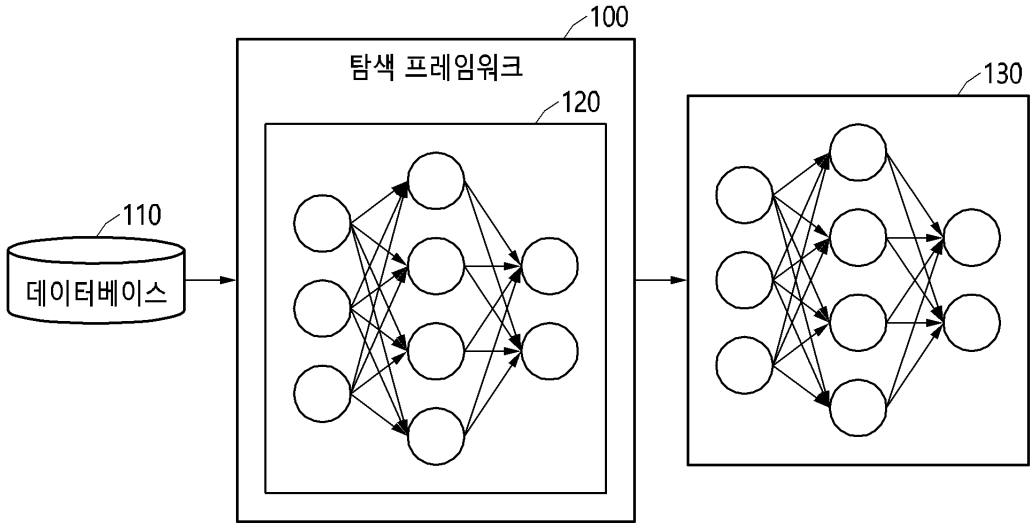
[0094] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0095] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

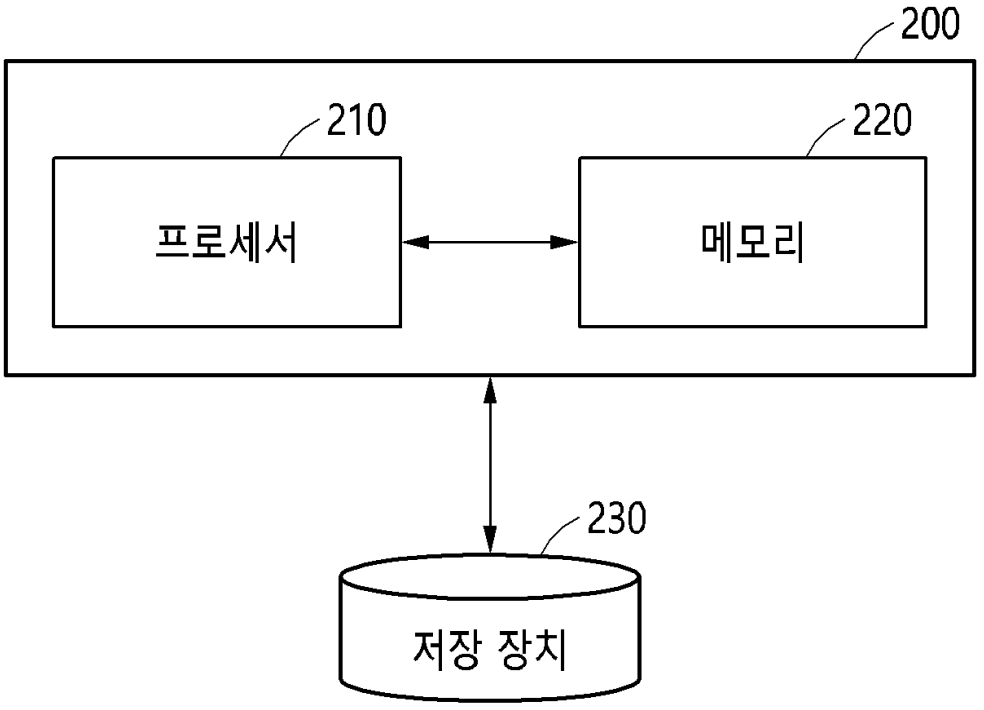
[0096] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

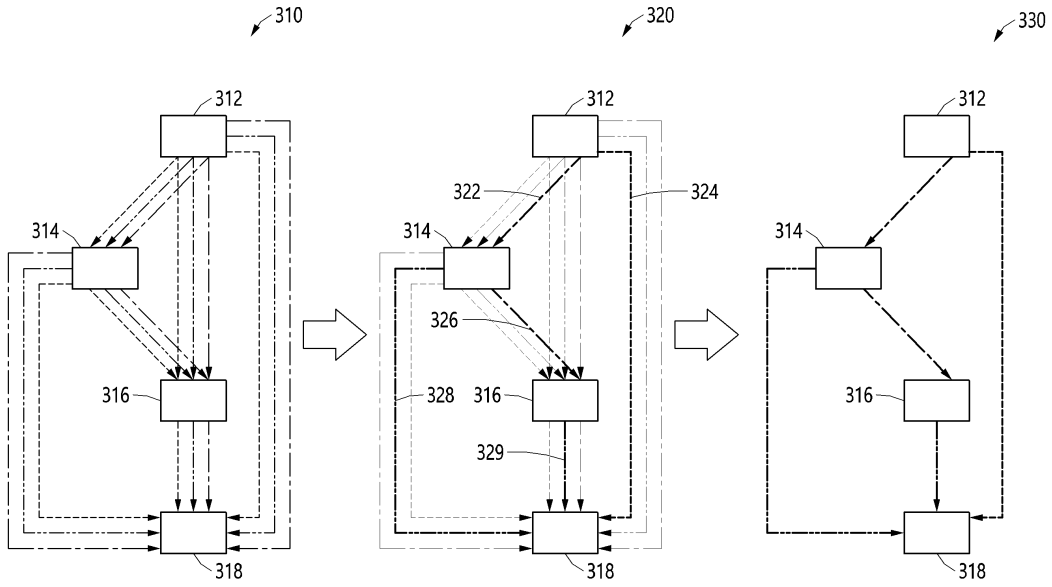
도면1



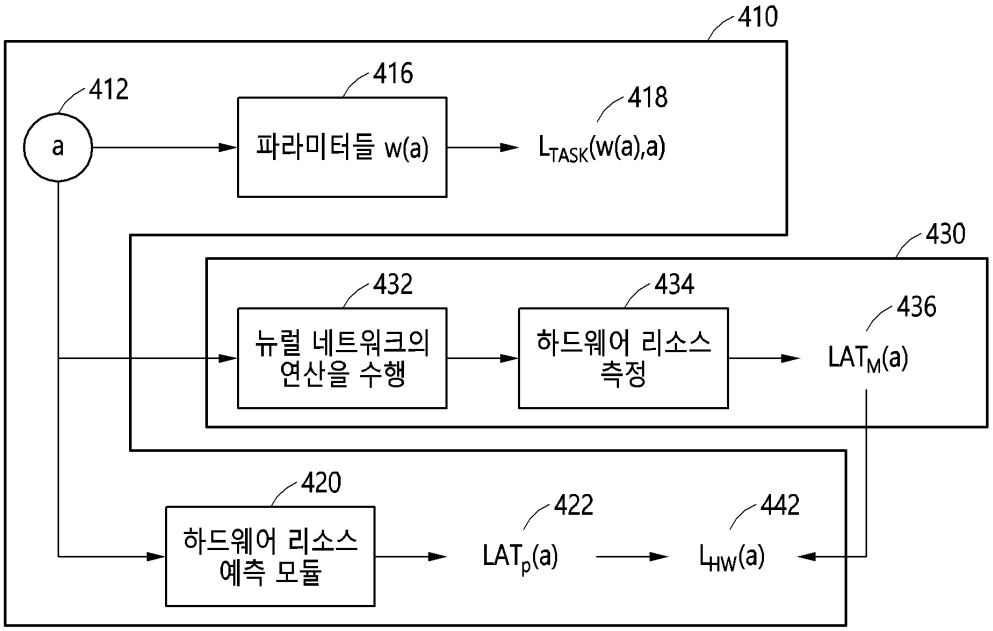
도면2



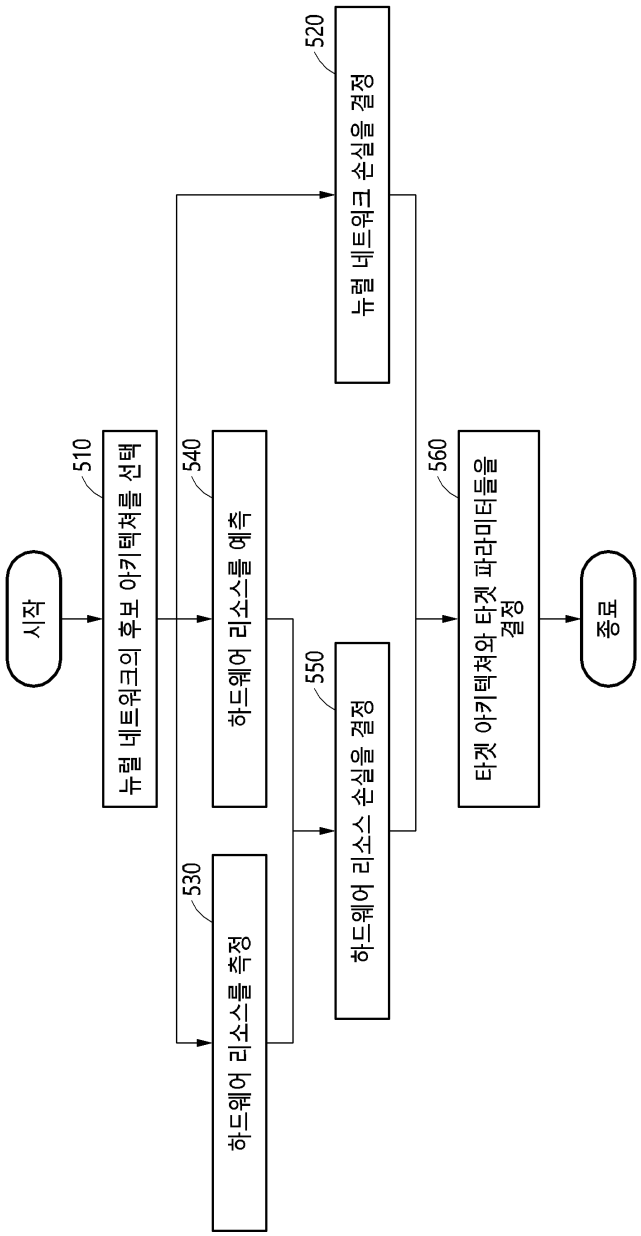
도면3


310
320
330
312
314
316
318
322
324
326
328
329

도면4


410
412
a
416
파라미터들 w(a)
418
$L_{TASK}(w(a),a)$
430
432
뉴럴 네트워크의 연산을 수행
434
하드웨어 리소스 측정
436
$LAT_M(a)$
420
하드웨어 리소스 예측 모듈
422
$LAT_p(a)$
442
$L_{HW}(a)$

도면5


시작
510
뉴럴 네트워크의 후보 아키텍쳐를 선택
520
뉴럴 네트워크 손실을 결정
530
하드웨어 리소스를 측정
540
하드웨어 리소스를 예측
550
하드웨어 리소스 손실을 결정
560
타겟 아키텍쳐와 타겟 파라미터들을 결정
종료

도면6