(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2017-0106338
(43) 공개일자 2017년09월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/04 (2006.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06N 3/04 (2013.01)
G06N 3/082 (2013.01)
(21) 출원번호 10-2017-7020008
(22) 출원일자(국제) 2015년12월15일
심사청구일자 없음
(85) 번역문제출일자 2017년07월18일
(86) 국제출원번호 PCT/US2015/065783
(87) 국제공개번호 WO 2016/118257
국제공개일자 2016년07월28일
(30) 우선권주장
62/106,608 2015년01월22일 미국(US)
14/846,579 2015년09월04일 미국(US)

(71) 출원인
퀄컴 인코포레이티드
미국 92121-1714 캘리포니아주 샌 디에고 모어하우스 드라이브 5775
(72) 발명자
안나푸레디 벤카타 스레칸타 레디
미국 92121-1714 캘리포니아주 샌디에고 모어하우스 드라이브 5775
디크만 다니엘 헨드리쿠스 프란시스쿠스
미국 92121-1714 캘리포니아주 샌디에고 모어하우스 드라이브 5775
줄리안 데이비드 조나단
미국 92121-1714 캘리포니아주 샌디에고 모어하우스 드라이브 5775
(74) 대리인
특허법인코리아나

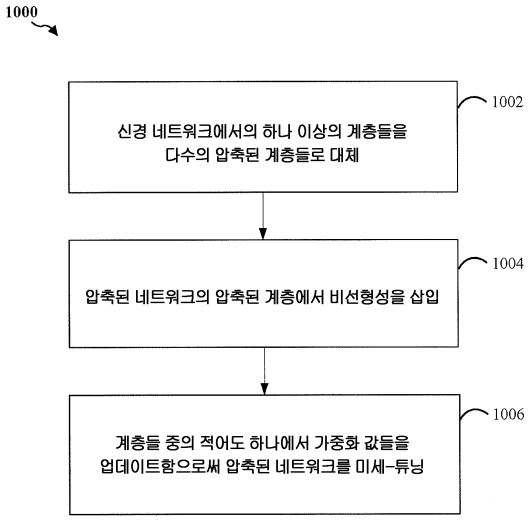
전체 청구항 수 : 총 36 항

(54) 발명의 명칭 **모델 압축 및 미세-튜닝**

(57) 요 약

신경 네트워크와 같은 머신 학습 네트워크를 압축하는 것은 압축된 네트워크를 생성하기 위하여 신경 네트워크에서의 하나의 계층을 압축된 계층들로 대체하는 것을 포함한다. 압축된 네트워크는 압축된 계층 (들) 에서 가중치 값들을 업데이트함으로서 미세-튜닝될 수도 있다.

대 표 도 - 도10



명 세 서

청구범위

청구항 1

신경 네트워크를 압축하는 방법으로서,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하는 단계;

압축된 상기 네트워크의 압축된 계층들 사이에 비선형성을 삽입하는 단계; 및

상기 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 2

제 1 항에 있어서,

상기 비선형성을 삽입하는 단계는 비선형 활성화 함수를 상기 압축된 계층들의 뉴런들에 적용하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 3

제 2 항에 있어서,

상기 비선형 활성화 함수는 교정기, 절대 값 함수, 쌍곡선 탄젠트 함수, 또는 시그모이드 함수 (sigmoid function) 인, 신경 네트워크를 압축하는 방법.

청구항 4

제 1 항에 있어서,

상기 미세-튜닝하는 단계는 상기 압축된 신경 네트워크에서 상기 가중치 값들을 업데이트함으로써 수행되는, 신경 네트워크를 압축하는 방법.

청구항 5

제 4 항에 있어서,

상기 미세-튜닝하는 단계는 비압축된 계층들의 서브세트에서 또는 상기 압축된 계층들의 서브세트의 적어도 하나에서 가중치 값들을 업데이트하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 6

제 4 항에 있어서,

상기 미세-튜닝하는 단계는 훈련 예들을 이용하여 수행되고, 상기 훈련 예들은 비압축된 네트워크를 훈련하기 위하여 이용된 예들의 제 1 세트, 또는 예들의 새로운 세트 중의 적어도 하나를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 7

제 1 항에 있어서,

압축을 반복적으로 적용하는 것, 비선형 계층들의 삽입, 및 더욱 심층 신경 네트워크들을 초기화하기 위한 방법으로서의 상기 미세-튜닝에 의해 상기 신경 네트워크를 초기화하는 단계를 더 포함하는, 신경 네트워크를 압축하는 방법.

청구항 8

신경 네트워크를 압축하는 방법으로서,

압축된 신경 네트워크를 생성하기 위하여, 상기 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 상기 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하는 단계; 및

적어도 하나의 압축된 계층에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 9

제 8 항에 있어서,

비압축된 계층들의 커널 크기는 상기 수용 필드 크기와 동일한, 신경 네트워크를 압축하는 방법.

청구항 10

제 8 항에 있어서,

상기 대체하는 것은, 특성들 $(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$ 및 $(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$ 이 충족되는 압축된 상기 네트워크를 생성하기 위하여, 커널 크기 $k_x \times k_y$ 를 가지는 상기 신경 네트워크에서의 적어도 하나의 계층을 커널 크기들 $k_{1x} \times k_{1y}, k_{2x} \times k_{2y} \ldots k_{\mathrm{L}x} \times k_{\mathrm{L}y}$ 을 갖는 동일한 타입의 다수의 압축된 계층들로 대체하는 것을 포함하는, 신경 네트워크를 압축하는 방법.

청구항 11

제 10 항에 있어서,

상기 커널 크기 $k_x \times k_y$ 를 갖는 컨볼루션 계층은 상기 커널 크기들 1 x 1, $k_x \times k_y$, 및 1 x 1 을 각각 갖는 3 개의 컨볼루션 계층들로 대체되는, 신경 네트워크를 압축하는 방법.

청구항 12

신경 네트워크를 압축하는 방법으로서,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하는 단계; 및

교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 13

제 12 항에 있어서,

상기 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하는 단계를 더 포함하는, 신경 네트워크를 압축하는 방법.

청구항 14

제 13 항에 있어서,

상기 미세-튜닝하는 단계는 비압축된 계층들의 서브세트에서 또는 상기 압축된 계층들의 서브세트의 적어도 하나에서 가중치 값들을 업데이트하는 단계를 포함하는, 신경 네트워크를 압축하는 방법.

청구항 15

제 13 항에 있어서,

상기 미세-튜닝하는 단계는, 제 1 스테이지에서, 상기 미세-튜닝하는 단계가 압축된 계층들의 서브세트에 대해 수행되고, 제 2 스테이지에서, 상기 미세-튜닝하는 단계가 압축 및 비압축된 계층들의 서브세트에 대해 수행되는 다수의 스테이지들에서 수행되는, 신경 네트워크를 압축하는 방법.

청구항 16

신경 네트워크를 압축하기 위한 장치로서,

메모리; 및

상기 메모리에 결합된 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하고;

압축된 상기 네트워크의 압축된 계층들 사이에 비선형성을 삽입하고; 그리고

적어도 하나의 압축된 계층들에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하도록 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 17

제 16 항에 있어서,

상기 적어도 하나의 프로세서는 비선형 활성화 함수를 상기 압축된 계층들의 뉴런들에 적용함으로써 비선형성을 삽입하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 18

제 17 항에 있어서,

상기 비선형 활성화 함수는 교정기, 절대 값 함수, 쌍곡선 탄젠트 함수, 또는 시그모이드 함수인, 신경 네트워크를 압축하기 위한 장치.

청구항 19

제 16 항에 있어서,

상기 적어도 하나의 프로세서는 상기 압축된 신경 네트워크에서 상기 가중치 값들을 업데이트함으로써 상기 미세-튜닝을 수행하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 20

제 19 항에 있어서,

상기 적어도 하나의 프로세서는 비압축된 계층들의 서브세트에서 또는 압축된 계층들의 서브세트의 적어도 하나에서 가중치 값들을 업데이트함으로써 상기 미세-튜닝을 수행하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 21

제 19 항에 있어서,

상기 적어도 하나의 프로세서는 훈련 예들을 이용함으로써 상기 미세-튜닝을 수행하도록 추가로 구성되고, 상기 훈련 예들은 비압축된 네트워크를 훈련하기 위하여 이용된 예들의 제 1 세트, 또는 예들의 새로운 세트 중의 적어도 하나를 포함하는, 신경 네트워크를 압축하기 위한 장치.

청구항 22

제 16 항에 있어서,

상기 적어도 하나의 프로세서는 압축을 반복적으로 적용하는 것, 비선형 계층들의 삽입, 및 더욱 심층 신경 네트워크들을 초기화하기 위한 방법으로서의 상기 미세-튜닝에 의해 상기 신경 네트워크를 초기화하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 23

신경 네트워크를 압축하기 위한 장치로서,

메모리; 및

상기 메모리에 결합된 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

압축된 신경 네트워크를 생성하기 위하여, 상기 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 상기 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드 크기와 일치하도록 하고; 그리고

압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하도록 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 24

제 23 항에 있어서,

비압축된 계층들의 커널 크기는 상기 수용 필드 크기와 동일한, 신경 네트워크를 압축하기 위한 장치.

청구항 25

제 23 항에 있어서,

상기 적어도 하나의 프로세서는 특성들 $(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$ 및 $(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$ 이 충족되는 압축된 상기 네트워크를 생성하기 위하여, 커널 크기 $k_x \times k_y$ 를 가지는 상기 신경 네트워크에서의 적어도 하나의 계층을 커널 크기들 $k_{1x} \times k_{1y}, k_{2x} \times k_{2y} \ldots k_{\mathrm{L}x} \times k_{\mathrm{L}y}$ 을 갖는 동일한 타입의 다수의 압축된 계층들로 대체하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 26

제 25 항에 있어서,

상기 커널 크기 $k_x \times k_y$ 를 갖는 컨볼루션 계층은 상기 커널 크기들 1 x 1, $k_x \times k_y$, 및 1 x 1 을 각각 갖는 3 개의 컨볼루션 계층들로 대체되는, 신경 네트워크를 압축하기 위한 장치.

청구항 27

신경 네트워크를 압축하기 위한 장치로서,

메모리; 및

상기 메모리에 결합된 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층

들로 대체하고; 그리고

교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하도록 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 28

제 27 항에 있어서,

상기 적어도 하나의 프로세서는 상기 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 29

제 28 항에 있어서,

상기 적어도 하나의 프로세서는 비압축된 계층들의 서브세트에서 또는 상기 압축된 계층들의 서브세트의 적어도 하나에서 가중치 값들을 업데이트함으로써 상기 미세-튜닝을 수행하도록 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 30

제 28 항에 있어서,

상기 적어도 하나의 프로세서는 제 1 스테이지에서, 상기 미세-튜닝이 압축된 계층들의 서브세트에 대해 수행되고, 제 2 스테이지에서, 상기 미세-튜닝이 압축 및 비압축된 계층들의 서브세트에 대해 수행되는 다수의 스테이지들에서 상기 미세-튜닝을 수행하도록 추가로 구성되는, 신경 네트워크를 압축하기 위한 장치.

청구항 31

신경 네트워크를 압축하기 위한 장치로서,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하기 위한 수단;

압축된 상기 네트워크의 압축된 계층들 사이에 비선형성을 삽입하기 위한 수단; 및

상기 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하기 위한 수단을 포함하는, 신경 네트워크를 압축하기 위한 장치.

청구항 32

신경 네트워크를 압축하기 위한 장치로서,

압축된 신경 네트워크를 생성하기 위하여, 상기 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 상기 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하기 위한 수단; 및

적어도 하나의 압축된 계층에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하기 위한 수단을 포함하는, 신경 네트워크를 압축하기 위한 장치.

청구항 33

신경 네트워크를 압축하기 위한 장치로서,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하기 위한 수단; 및

교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하기 위한 수단을 포함하는, 신경 네트워크를 압축하기 위한 장치.

청구항 34

신경 네트워크를 압축하기 위한 프로그램 코드가 인코딩된 비-일시적 컴퓨터 판독가능 저장 매체로서,

상기 프로그램 코드는 프로세서에 의해 실행되고, 상기 프로그램 코드는

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하기 위한 프로그램 코드;

압축된 상기 네트워크의 압축된 계층들 사이에 비선형성을 삽입하기 위한 프로그램 코드; 및

적어도 하나의 압축된 계층들에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하기 위한 프로그램 코드를 포함하는, 비-일시적 컴퓨터 판독가능 저장 매체.

청구항 35

신경 네트워크를 압축하기 위한 프로그램 코드가 인코딩된 비-일시적 컴퓨터 판독가능 저장 매체로서,

상기 프로그램 코드는 프로세서에 의해 실행되고,

압축된 신경 네트워크를 생성하기 위하여, 상기 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 상기 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드 크기와 일치하도록 하기 위한 프로그램 코드; 및

압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 상기 네트워크를 미세-튜닝하기 위한 프로그램 코드를 포함하는, 비-일시적 컴퓨터 판독가능 저장 매체.

청구항 36

신경 네트워크를 압축하기 위한 프로그램 코드가 인코딩된 비-일시적 컴퓨터 판독가능 저장 매체로서,

상기 프로그램 코드는 프로세서에 의해 실행되고,

압축된 신경 네트워크를 생성하기 위하여 상기 신경 네트워크에서의 적어도 하나의 계층을 복수의 압축된 계층들로 대체하기 위한 프로그램 코드; 및

교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하기 위한 프로그램 코드를 포함하는, 비-일시적 컴퓨터 판독가능 저장 매체.

발명의 설명

기 술 분 야

[0001] **관련 출원에 대한 상호-참조**

[0002] 본 출원은 "MODEL COMPRESSION AND FINE-TUNING" 이라는 명칭인, 2015 년 1 월 22 일자로 출원된 미국 특허 가출원 제 62/106,608 호의 이익을 주장하고, 그 개시물은 그 전체적으로 참조로 본원에 명백히 편입된다.

[0003] 본 개시물의 어떤 양태들은 일반적으로, 신경 시스템 공학에 관한 것으로, 더욱 상세하게는, 신경 네트워크를 압축하기 위한 시스템들 및 방법들에 관한 것이다.

배 경 기 술

[0004] 인공 뉴런 (artificial neuron) 들 (예컨대, 뉴런 모델들) 의 상호접속된 그룹을 포함할 수도 있는 인공 신경 네트워크는 연산 디바이스이거나, 연산 디바이스에 의해 수행되어야 할 방법을 표현한다.

[0005] 컨볼루션 신경 네트워크 (convolutional neural network) 들은 피드-포워드 인공 신경 네트워크 (feed-forward artificial neural network) 의 타입이다. 컨볼루션 신경 네트워크들은, 수용 필드 (receptive field) 를 각각 가지며 입력 공간을 집합적으로 타일링 (tiling) 하는 뉴런들의 집합들을 포함할 수도 있다. 컨볼루션 신경 네트워크 (convolutional neural network; CNN) 들은 많은 애플리케이션들을 가진다. 특히, CNN 들은 패턴 인식 및 분류의 영역에서 폭넓게 이용되었다.

[0006] 심층 신뢰 네트워크 (deep belief network) 들 및 심층 컨볼루션 네트워크 (deep convolutional network) 들과 같은 심층 학습 (deep learning) 아키텍처들은, 뉴런들의 제 1 계층의 출력이 뉴런들의 제 2 계층에 대한 입력

이 되고, 뉴런들의 제 2 계층의 출력이 뉴런들의 제 3 계층에 대한 입력이 되는 등등과 같은 계층화된 신경 네트워크들 아키텍처들이다. 심층 신경 네트워크들은 특징들의 계층구조를 인식하도록 훈련될 수도 있고, 따라서, 그것들은 객체 인식 애플리케이션들에서 점점 더 이용되었다. 컨볼루션 심경 네트워크들과 같이, 이 심층 학습 아키텍처들에서의 연산은 하나 이상의 연산 체인 (computational chain) 들로 구성될 수도 있는 프로세싱 노드들의 집단 상에서 분산될 수도 있다. 이 멀티-계층화된 아키텍처들은 한 번에 하나의 계층에서 훈련될 수도 있고, 역 전파 (back propagation) 를 이용하여 미세-튜닝될 수도 있다.

[0007] 다른 모델들은 또한, 객체 인식을 위하여 이용가능하다. 예를 들어, 지원 벡터 머신 (support vector machine; SVM) 들은 분류를 위하여 적용될 수 있는 학습 도구들이다. 지원 벡터 머신들은 데이터를 범주화하는 분리 초평면 (separating hyperplane) (예컨대, 판단 경계) 을 포함한다. 초평면은 감독된 학습에 의해 정의된다. 바람직한 초평면은 훈련 데이터의 마진을 증가시킨다. 다시 말해서, 초평면은 훈련 예들까지의 가장 큰 최소 거리를 가져야 한다.

[0008] 이 해결책들은 다수의 분류 벤치마크 (benchmark) 들에 대한 우수한 결과들을 달성하지만, 그 연산 복잡도는 엄청나게 높을 수 있다. 추가적으로, 모델들의 훈련은 도전일 수도 있다.

발명의 내용

[0009] 하나의 양태에서는, 신경 네트워크와 같은 머신 학습 네트워크를 압축하는 방법이 개시되어 있다. 방법은 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하는 단계를 포함한다. 방법은 또한, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하는 단계를 포함한다. 또한, 방법은 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하는 단계를 포함한다.

[0010] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하기 위한 수단을 포함한다. 장치는 또한, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하기 위한 수단을 포함한다. 또한, 장치는 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하기 위한 수단을 포함한다.

[0011] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 메모리, 및 메모리에 결합된 적어도 하나의 프로세서를 포함한다. 프로세서 (들) 는 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하도록 구성된다. 프로세서 (들) 는 또한, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하도록 구성된다. 또한, 프로세서 (들) 는 또한, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하도록 구성된다.

[0012] 또 다른 양태는 비-일시적 컴퓨터 판독가능 매체를 개시한다. 컴퓨터 판독가능 매체는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위하여 비-일시적 프로그램 코드가 기록된다. 프로그램 코드는, 프로세서 (들) 에 의해 실행될 때, 프로세서 (들) 로 하여금, 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하게 한다. 프로그램 코드는 또한, 프로세서 (들) 로 하여금, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하게 한다. 또한, 프로그램 코드는 또한, 프로세서 (들) 로 하여금, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하게 한다.

[0013] 또 다른 양태에서는, 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 방법이 개시되어 있다. 방법은 압축된 신경 네트워크를 생성하기 위하여, 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하는 단계를 포함한다. 방법은 또한, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하는 단계를 포함한다.

[0014] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 압축된 신경 네트워크를 생성하기 위하여, 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하기 위한 수단을 포함한다. 장치는 또한, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트

워크를 미세-튜닝하기 위한 수단을 포함한다.

[0015] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 메모리, 및 메모리에 결합된 적어도 하나의 프로세서를 포함한다. 프로세서 (들) 는 압축된 신경 네트워크를 생성하기 위하여, 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하게 하도록 구성된다. 프로세서 (들) 는 또한, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하도록 구성된다.

[0016] 또 다른 양태는 비-일시적 컴퓨터 판독가능 매체를 개시한다. 컴퓨터 판독가능 매체는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위하여 비-일시적 프로그램 코드가 기록된다. 프로그램 코드는, 프로세서 (들) 에 의해 실행될 때, 프로세서 (들) 로 하여금, 압축된 신경 네트워크를 생성하기 위하여, 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하게 한다. 프로그램 코드는 또한, 프로세서 (들) 로 하여금, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하게 한다.

[0017] 또 다른 양태에서는, 신경 네트워크와 같은 머신 학습 네트워크를 압축하는 방법이 개시되어 있다. 방법은 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하는 단계를 포함한다. 방법은 또한, 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스 (weight matrix) 들을 결정하는 단계를 포함한다.

[0018] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하기 위한 수단을 포함한다. 장치는 또한, 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하기 위한 수단을 포함한다.

[0019] 또 다른 양태는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위한 장치를 개시한다. 장치는 메모리, 및 메모리에 결합된 적어도 하나의 프로세서를 포함한다. 프로세서 (들) 는 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하도록 구성된다. 프로세서 (들) 는 또한, 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하도록 구성된다.

[0020] 또 다른 양태는 비-일시적 컴퓨터 판독가능 매체를 개시한다. 컴퓨터 판독가능 매체는 신경 네트워크와 같은 머신 학습 네트워크를 압축하기 위하여 프로그램 코드가 기록된다. 프로그램 코드는, 프로세서 (들) 에 의해 실행될 때, 프로세서 (들) 로 하여금, 압축된 신경 네트워크를 생성하기 위하여 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하게 한다. 프로그램 코드는 또한, 프로세서 (들) 로 하여금, 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하게 한다.

[0021] 개시물의 추가적인 특징들 및 장점들이 이하에서 설명될 것이다. 이 개시물은 본 개시물의 동일한 목적들을 수행하기 위한 다른 구조들을 수행하거나 설계하기 위한 기초로서 용이하게 사용될 수도 있다는 것이 당해 분야의 당업자들에 의해 인식되어야 한다. 또한, 이러한 등가적인 구성들은 첨부된 청구항들에서 기재된 바와 같은 개시물의 교시사항들로부터 이탈하지 않는다는 것이 당해 분야의 당업자들에 의해 실현되어야 한다. 그 편성 및 동작 방법의 양자에 대하여, 개시물의 특성인 것으로 믿어지는 신규한 특징들은 추가의 목적들 및 장점들과 함께, 동반된 도면들과 관련하여 고려될 때에 다음의 설명으로부터 더욱 양호하게 이해될 것이다. 그러나, 도면들의 각각은 본 개시물의 제한들의 정의로서 의도된 것이 아니라, 오직 예시 및 설명의 목적들을 위하여 제공된다는 것이 명백히 이해되어야 한다.

도면의 간단한 설명

[0022] 본 개시물의 특징들, 본질, 및 장점들은 유사한 참조 부호들이 이에 대응하여 전반에 걸쳐 식별하는 도면들과 함께 취해질 때에 이하에서 기재된 상세한 설명으로부터 더욱 분명해질 것이다.

도 1 은 본 개시물의 어떤 양태들에 따라, 범용 프로세서를 포함하는 시스템-온-어-칩 (system-on-a-chip; SOC) 을 이용하여 신경 네트워크를 설계하는 일 예의 구현예를 예시한다.

도 2 는 본 개시물의 양태들에 따라 시스템의 일 예의 구현예를 예시한다.

도 3a 는 본 개시물의 양태들에 따라 신경 네트워크를 예시하는 도면이다.

도 3b 는 본 개시물의 양태들에 따라 예시적인 심층 컨볼루션 네트워크 (deep convolutional network; DCN) 를 예시하는 블록도이다.

도 4a 는 본 개시물의 양태들에 따라 인공 지능 (artificial intelligence; AI) 기능들을 모듈화할 수도 있는 예시적인 소프트웨어 아키텍처를 예시하는 블록도이다.

도 4b 는 본 개시물의 양태들에 따라 스마트폰 상에서의 AI 애플리케이션의 실행 시간 (run-time) 동작을 예시하는 블록도이다.

도 5a 내지 도 5b 및 도 6a 내지 도 6b 는 본 개시물의 양태들에 따라 완전히-접속된 계층 및 압축된 완전히-접속된 계층을 예시하는 블록도들이다.

도 7 은 본 개시물의 양태들에 따라 예시적인 컨볼루션 계층을 예시하는 블록도이다.

도 8a 내지 도 8b 및 도 9a 내지 도 9b 는 본 개시물의 양태들에 따라 컨볼루션 계층의 일 예의 압축을 예시한다.

도 10 내지 도 13 은 본 개시물의 양태들에 따라 신경 네트워크를 압축하기 위한 방법들을 예시하는 흐름도들이다.

발명을 실시하기 위한 구체적인 내용

[0023] 이하에서 기재된 상세한 설명은 첨부된 도면들과 관련하여, 다양한 구성들의 설명으로서 의도되고, 본원에서 설명된 개념들이 실시될 수도 있는 유일한 구성들을 표현하도록 의도된 것이 아니다. 상세한 설명은 다양한 개념들의 철저한 이해를 제공하는 목적을 위한 특정 세부사항들을 포함한다. 그러나, 이 개념들은 이 특정 세부사항들 없이 실시될 수도 있다는 것이 당해 분야의 당업자들에게 명백할 것이다. 일부 사례들에서는, 이러한 개념들을 모호하게 하는 것을 회피하기 위하여, 잘 알려진 구조들 및 컴포넌트들이 블록도 형태로 도시되어 있다.

[0024] 교시사항들에 기초하여, 당해 분야의 당업자는 개시물의 범위가, 개시물의 임의의 다른 양태에 관계 없이 또는 이와 조합하여 구현되든지 간에, 개시물의 임의의 양태를 커버하도록 의도된다는 것을 인식해야 한다. 예를 들어, 기재된 임의의 수의 양태들을 이용하여 장치가 구현될 수도 있거나 방법이 실시될 수도 있다. 게다가, 개시물의 범위는 기재된 개시물의 다양한 양태들에 추가하여, 또는 이 다양한 양태들 이외에, 다른 구조, 기능성, 또는 구조 및 기능성을 이용하여 실시된 이러한 장치 또는 방법을 커버하도록 의도된다. 개시된 개시물의 임의의 양태는 청구항의 하나 이상의 구성요소들에 의해 구체화될 수도 있다는 것이 이해되어야 한다.

[0025] 단어 "예시적" 은 "예, 사례, 또는 예시로서 작용함" 을 의미하기 위하여 본원에서 이용된다. "예시적" 으로서 본원에서 설명된 임의의 양태는 다른 양태들에 비해 바람직하거나 유익한 것으로 반드시 해석되어야 하는 것은 아니다.

[0026] 특정한 양태들이 본원에서 설명되지만, 이 양태들의 많은 변형들 및 치환들은 개시물의 범위 내에 속한다. 바람직한 양태들의 일부 이득들 및 장점들이 언급되지만, 개시물의 범위는 특정한 이득들, 용도들, 또는 목적들에 제한되도록 의도된 것이 아니다. 오히려, 개시물의 양태들은 상이한 기술들, 시스템 구성들, 네트워크들, 및 프로토콜들에 폭넓게 적용가능하도록 의도되며, 이들의 일부는 바람직한 양태들의 도면들 및 다음의 설명에서 예로서 예시되어 있다. 상세한 설명 및 도면들은 제한하는 것이 아니라 개시물의 단지 예시이고, 개시물의 범위는 첨부된 청구항들 및 그 등가물들에 의해 정의된다.

[0027] 모델 압축 및 미세-튜닝

[0028] 심층 신경 네트워크들은 이미지/비디오 분류, 스피치 인식, 얼굴 인식 등과 같은 몇몇 인공 지능 태스크들에서 최신 기술을 수행한다. 신경 네트워크 모델들은 훈련 예들의 대형 데이터베이스로부터 훈련되고, 전형적으로, 더욱 대형의 모델들은 더욱 양호한 성능을 달성하는 경향이 있다. 스마트폰들, 로봇들, 및 자동차들과 같은 이동 디바이스들에 관한 이 신경 네트워크 모델들을 개발할 목적을 위해서는, 연산 복잡도, 메모리 풋프린트 (memory footprint), 및 전력 소비를 가능한 한 많이 감소시키거나 최소화하는 것이 바람직하다. 이것들은 또한, 하루 당 수백만 개의 이미지들/비디오들을 프로세싱하는 데이터 센터들과 같은 클라우드 애플리케이션

들을 위한 바람직한 특성들이다.

[0029] 본 개시물의 양태들은 신경 네트워크 모델들을 압축하는 방법에 관한 것이다. 압축된 모델은 상당히 더 적은 수의 파라미터들을 가지고, 이 때문에, 더 작은 메모리 풋프린트를 가진다. 게다가, 압축된 모델은 상당히 더 적은 동작들을 가지고, 이 때문에, 더욱 고속의 추론 실행 시간을 초래한다.

[0030] 컨볼루션 신경 네트워크 모델들은 계층들의 시퀀스로 분할될 수도 있다. 각각의 계층은 네트워크에서의 하나 이상의 선행 계층들로부터 수신된 입력을 변환할 수도 있고, 네트워크의 후속 계층들에 공급될 수도 있는 출력을 생성할 수도 있다. 예를 들어, 컨볼루션 신경 네트워크는 완전히-접속된 계층들, 컨볼루션 계층들, 국소적으로-접속된 계층들, 및 다른 계층들을 포함할 수도 있다. 상이한 계층들의 각각은 상이한 타입의 변환을 수행할 수도 있다.

[0031] 본 개시물의 양태들에 따르면, 완전히-접속된 계층들, 컨볼루션 계층들, 및 국소적으로-접속된 계층들이 압축될 수도 있다. 이 계층들의 전부는 입력 값들에 관한 선형 변환을 적용할 수도 있고, 여기서, 선형 변환들은 가중치 값들에 의해 파라미터화된다. 가중치 값들은 2 차원 행렬들 또는 더 높은 차원의 텐서 (tensor) 들로서 표현될 수도 있다. 선형 변환 구조는 상이한 계층 타입들에 대하여 상이할 수도 있다.

[0032] 일부 양태들에서, 계층은 그것을 동일한 타입의 다수의 계층들로 대체함으로써 압축될 수도 있다. 예를 들어, 컨볼루션 계층은 그것을 다수의 컨볼루션 계층들로 대체함으로써 압축될 수도 있고, 완전히-접속된 계층은 그것을 다수의 완전히-접속된 계층들로 대체함으로써 압축될 수도 있다. 압축된 계층에 있는 뉴런들은 아이덴티티 활성화 함수 (identity activation function) 로 구성될 수도 있다. 하나의 계층은 다수의 압축된 계층들로 대체될 수도 있지만, 모든 압축된 계층들 모두의 합계 복잡도 (sum complexity) (및 메모리 풋프린트) 는 단일의 비압축된 계층의 그것보다 더 작을 수도 있다.

[0033] 그러나, 모델 압축은 신경 네트워크의 성능에 있어서의 하락을 희생하여 나올 수도 있다. 예를 들어, 신경 네트워크가 분류 문제를 위하여 이용될 경우, 분류 정확도는 압축으로 인해 하락할 수도 있다. 더 높은 압축은 정확도에 있어서의 더 높은 하락으로 귀착되므로, 정확도에 있어서의 이 하락은 압축의 정도를 제한한다.

정확도에 있어서의 하락을 방지하기 위하여, 일부 양태들에서, 압축된 모델은 훈련 예들을 이용하여 미세-튜닝될 수도 있다. 압축된 모델들의 미세-튜닝은 정확도에 있어서의 하락을 되찾을 수도 있고, 이에 따라, 분류 정확도에 있어서의 너무 많은 하락 없는 개선된 압축으로 귀착될 수도 있다. 압축된 모델의 미세-튜닝은 계층-대-계층 (layer-by-layer) 에 기초하여 행해질 수도 있다.

[0034] 이에 따라, 본 개시물의 양태들은 신경 네트워크의 계층이 더 많은 이러한 계층들로 대체되는 압축 기법들을 제공한다. 이 때문에, 결과적인 압축된 모델은 또 다른 컨볼루션 신경 네트워크 모델이다. 이것은 압축된 모델을 구현하기 위한 새롭고 명시적인 기법들의 개발이 회피될 수도 있기 때문에 상당한 장점이다. 즉, 임의의 신경 네트워크 라이브러리가 원래의 비압축된 모델과 호환가능할 경우, 그것은 또한 압축된 모델과 호환가능할 것이다. 또한, 결과적인 압축된 모델은 컨볼루션 신경 네트워크이므로, 미세-튜닝을 특정한 계층으로 제한하기 보다는, 전체 스택 미세-튜닝 (예컨대, 역 전파) 이 수행될 수도 있다.

[0035] 신경 네트워크는 몇몇 계층들로 이루어진다. 각각의 계층은 하나 이상의 이전의 계층들로부터의 활성화 벡터들을 입력으로서 취하고, 조합된 입력 벡터에 관한 선형/비선형 변환을 적용하고, 후속 계층들에 의해 이용되어야 할 활성화 벡터를 출력한다. 일부 계층들은 가중치 값들로 파라미터화되는 반면, 일부 계층들은 그렇지 않다. 본 개시물의 양태들은 가중화 계층들과 관련된다: 1) 완전히-접속된 계층들, 2) 컨볼루션 계층들, 및 3) 국소적으로-접속된 계층들. 모든 3 개의 계층들은 선형 변환을 수행하지만, 출력 뉴런들이 어떻게 입력 뉴런들에 접속되는지에 있어서 상이하다.

[0036] 도 1 은 본 개시물의 어떤 양태들에 따라 범용 프로세스 (CPU) 또는 멀티-코어 범용 프로세서들 (CPU 들) (102) 을 포함할 수도 있는 시스템-온-어-칩 (SOC) (100) 을 이용하여 상기 언급된 신경 네트워크를 압축하는 일 예의 구현예를 예시한다. 변수 (예컨대, 신경 신호 및 시냅틱 가중치 (synaptic weight)), 연산 디바이스 (예컨대, 가중치들을 갖는 신경 네트워크) 와 연관된 시스템 파라미터들, 지연들, 주파수 빈 (frequency bin) 정보, 및 태스크 정보는 신경 프로세싱 유닛 (neural processing unit; NPU) (108) 와 연관된 메모리 블록에서, CPU (102) 와 연관된 메모리 블록에서, 그래픽 프로세싱 유닛 (graphics processing unit; GPU) (104) 과 연관된 메모리 블록에서, 디지털 신호 프로세서 (digital signal processor; DSP) (106) 와 연관된 메모리 블록에서, 전용 메모리 블록 (118) 에서 저장될 수도 있거나, 다수의 블록들에 걸쳐 분산될 수도 있다. 범용 프로세서 (102) 에서 실행된 명령들은 CPU (102) 와 연관된 프로그램 메모리로부터 로딩될 수도 있거나, 전용 메모리 블

록 (118) 으로부터 로딩될 수도 있다.

[0037] SOC (100) 는 또한, GPU (104), DSP (106), 4 세대 롱텀 에볼루션 (fourth generation long term evolution; 4G LTE) 접속성, 비인가된 Wi-Fi 접속성, USB 접속성, 블루투스 접속성 등등을 포함할 수도 있는 접속성 블록 (110), 및 예를 들어, 제스처 (gesture) 들을 검출할 수도 있고 인식할 수도 있는 멀티미디어 프로세서 (112) 와 같은, 특정 기능들에 맞추어진 추가적인 프로세싱 블록들을 포함할 수도 있다. 하나의 구현예에서, NPU 는 CPU, DSP, 및/또는 GPU 에서 구현된다. SOC (100) 는 또한, 센서 프로세서 (114), 이미지 신호 프로세서 (image signal processor; ISP) 들, 및/또는 글로벌 위치확인 시스템을 포함할 수도 있는 내비게이션 (120) 을 포함할 수도 있다.

[0038] SOC 는 ARM 명령 세트에 기초할 수도 있다. 본 개시물의 양태에서, 범용 프로세서 (102) 로 로딩된 명령들은, 압축된 네트워크를 생성하기 위하여, 신경 네트워크와 같은 머신 학습 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하고, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하고, 및/또는 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하기 위한 코드를 포함할 수도 있다.

[0039] 본 개시물의 또 다른 양태에서, 범용 프로세서 (102) 로 로딩된 명령들은, 압축된 네트워크를 생성하기 위하여, 신경 네트워크와 같은 머신 학습 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하게 하기 위한 코드를 포함할 수도 있다. 또한, 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하기 위한 코드가 있을 수도 있다.

[0040] 본 개시물의 또 다른 양태에서, 범용 프로세서 (102) 로 로딩된 명령들은, 압축된 네트워크를 생성하기 위하여, 신경 네트워크와 같은 머신 학습 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하기 위한 코드를 포함할 수도 있다. 또한, 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하기 위한 코드가 있을 수도 있다.

[0041] 도 2 는 본 개시물의 어떤 양태들에 따라 시스템 (200) 의 일 예의 구현예를 예시한다. 도 2 에서 예시된 바와 같이, 시스템 (200) 은 본원에서 설명된 방법들의 다양한 동작들을 수행할 수도 있는 다수의 로컬 프로세싱 유닛들 (202) 을 가질 수도 있다. 각각의 로컬 프로세싱 유닛 (202) 은 로컬 상태 메모리 (204), 및 신경 네트워크의 파라미터들을 저장할 수도 있는 로컬 파라미터 메모리 (206) 를 포함할 수도 있다. 게다가, 로컬 프로세싱 유닛 (202) 은 로컬 모델 프로그램을 저장하기 위한 로컬 (뉴런) 모델 프로그램 (local model program; LMP) 메모리 (208), 로컬 학습 프로그램을 저장하기 위한 로컬 학습 프로그램 (local learning program; LLP) 메모리 (210), 및 로컬 접속 메모리 (212) 를 가질 수도 있다. 또한, 도 2 에서 예시된 바와 같이, 각각의 로컬 프로세싱 유닛 (202) 은 로컬 프로세싱 유닛의 로컬 메모리들을 위한 구성들을 제공하기 위한 구성 프로세서 유닛 (214) 과 인터페이싱할 수도 있고, 로컬 프로세싱 유닛들 (202) 사이의 라우팅을 제공하는 라우팅 접속 프로세싱 유닛 (216) 과 인터페이싱할 수도 있다.

[0042] 심층 학습 아키텍처들은 각각의 계층에서의 추상화의 연속적으로 더 높은 레벨들에서 입력들을 표현하는 것을 학습함으로써 객체 인식 태스크를 수행할 수도 있고, 이것에 의해, 입력 데이터의 유용한 특징 표현을 구축할 수도 있다. 이러한 방법으로, 심층 학습은 전통적인 머신 학습의 주요한 병목 현상을 해결한다. 심층 학습의 출현 이전에, 객체 인식 문제에 대한 머신 학습 접근법은 아마도 얕은 분류기 (shallow classifier) 와 조합하여, 인간 공학적 특징들에 과도하게 의존하였을 수도 있다. 얕은 분류기는 입력이 어느 클래스에 속하는지를 예측하기 위하여, 특징 벡터 컴포넌트들의 가중화된 합계가 임계치와 비교될 수도 있는 예를 들어, 2-클래스 선형 분류기일 수도 있다. 인간 공학적 특징들은 도메인 전문지식을 갖는 공학자들에 의해 특정 문제 도메인에 맞추어진 템플릿 (template) 들 또는 커널 (kernel) 들일 수도 있다. 대조적으로, 심층 학습 아키텍처들은, 인간 공학자가 설계한 것과 유사한 특징들을 표현하도록 학습할 수도 있지만, 훈련을 통해 학습할 수도 있다. 또한, 심층 네트워크는 인간이 고려하지 않았을 수도 있는 새로운 타입들의 특징들을 표현하고 인식하는 것을 학습할 수도 있다.

[0043] 심층 학습 아키텍처는 특징들의 계층구조를 학습할 수도 있다. 예를 들어, 시각적 데이터가 제시될 경우, 제 1 계층은 입력 스트림에서 에지 (edge) 들과 같은 간단한 특징들을 인식하는 것을 학습할 수도 있다. 청각적 데이터가 제시될 경우, 제 1 계층은 특정 주파수들에서의 스펙트럼 파워 (spectral power) 를 인식하는 것을 학습할 수도 있다. 제 1 계층의 출력을 입력으로서 취하는 제 2 계층은 시각적 데이터를 위한 간단한 형상들 또는 청각적 데이터를 위한 사운드들의 조합들과 같은 특징들의 조합들을 인식하는 것을 학습할 수도

있다. 더 높은 계층들은 시각적 데이터에서의 복잡한 형상들 또는 청각적 데이터에서의 단어들을 표현하는 것을 학습할 수도 있다. 더욱 더 높은 계층들은 공통의 시각적 객체들 또는 발화된 어구 (spoken phrase) 들을 인식하는 것을 학습할 수도 있다.

[0044] 심층 학습 아키텍처들은 자연 계층적 구조를 가지는 문제들에 적용될 때에 특히 양호하게 수행할 수도 있다. 예를 들어, 자동차의 분류는 휠 (wheel), 윈드실드 (windshield), 및 다른 특징들을 인식하기 위한 첫번째 학습으로부터 이익을 얻을 수도 있다. 이 특징들은 승용차, 트럭, 및 비행기를 인식하기 위하여 상이한 방법들로 더 높은 계층들에서 조합될 수도 있다.

[0045] 신경 네트워크들은 다양한 접속성 패턴들로 설계될 수도 있다. 피드-포워드 네트워크들에서, 정보는 더 낮은 것으로부터 더 높은 계층들로 전달되고, 주어진 계층에 있는 각각의 뉴런은 더 높은 계층들에 있는 뉴런들로 통신한다. 계층적 표현은 위에서 설명된 바와 같이, 피드-포워드 네트워크의 연속적인 계층들에서 구축될 수도 있다. 신경 네트워크들은 또한, 회귀적 또는 피드백 (또한 하향식 (top-down) 으로 칭해짐) 접속들을 가질 수도 있다. 회귀적 접속에서, 주어진 계층에 있는 뉴런으로부터의 출력은 동일한 계층에 있는 또 다른 뉴런으로 통신된다. 회귀적 아키텍처는 시퀀스로 신경 네트워크에 전달되는 하나보다 많은 입력 데이터 청크 (chunk) 들에 걸치는 패턴들을 인식하는데 도움이 될 수도 있다. 주어진 계층에 있는 뉴런으로부터 더 낮은 계층에 있는 뉴런으로의 접속은 피드백 (또는 하향식) 접속으로 칭해진다. 하이-레벨 (high-level) 개념의 인식이 입력의 구체적인 로우-레벨 (low-level) 특징들을 구별하는 것을 도울 수도 있을 때에 다수의 피드백 접속들을 갖는 네트워크가 도움이 될 수도 있다.

[0046] 도 3a 를 참조하면, 신경 네트워크의 계층들 사이의 접속들은 완전히-접속 (302) 될 수도 있거나, 국소적으로 접속 (304) 될 수도 있다. 완전히-접속된 네트워크 (302) 에서, 제 1 계층에 있는 뉴런은 그 출력을 제 2 계층에 있는 매 뉴런으로 통신할 수도 있어서, 제 2 계층에 있는 각각의 뉴런은 제 1 계층에 있는 매 뉴런으로부터 입력을 수신할 것이다. 대안적으로, 국소적으로 접속된 네트워크 (304) 에서, 제 1 계층에 있는 뉴런은 제 2 계층에 있는 제한된 수의 뉴런들에 접속될 수도 있다. 컨볼루션 네트워크 (306) 는 국소적으로 접속될 수도 있고, 제 2 계층에 있는 각각의 뉴런에 대한 입력들과 연관된 접속 강도들이 공유 (예컨대, 308) 되도록 추가로 구성된다. 더욱 일반적으로, 네트워크의 국소적으로 접속된 계층은 계층에 있는 각각의 뉴런이 동일하거나 유사한 접속성 패턴을 가지도록, 그러나, 상이한 값들을 가질 수도 있는 접속들 강도들 (예컨대, 310, 312, 314, 및 316) 로 구성될 수도 있다. 주어진 영역에 있는 더 높은 계층 뉴런들이 네트워크에 대한 총 입력의 한정된 부분의 특성들에 대한 훈련을 통해 튜닝되는 입력들을 수신할 수도 있으므로, 국소적으로 접속된 접속성 패턴은 더 높은 계층에서의 공간적으로 별개의 수용적 필드들을 야기시킬 수도 있다.

[0047] 국소적으로 접속된 신경 네트워크들은 입력들의 공간적 로케이션이 의미 있는 문제들에 잘 맞을 수도 있다. 예를 들어, 차재 카메라로부터 시각적 특징들을 인식하도록 설계된 네트워크 (300) 는 이미지의 하부 대 상부 부분과의 그 연관성에 따라 상이한 특성들을 갖는 높은 계층 뉴런들을 개발할 수도 있다. 이미지의 하부 부분과 연관된 뉴런들은 예를 들어, 차선 표기들을 인식하는 것을 학습할 수도 있는 반면, 이미지의 하부 부분과 연관된 뉴런들은 신호등들, 교통 표지들 등등을 인식하는 것을 학습할 수도 있다.

[0048] DCN 은 감독된 학습으로 훈련될 수도 있다. 훈련하는 동안, DCN 은 속도 제한 표지의 크롭핑된 (cropped) 이미지와 같은 이미지 (326) 를 제시받을 수도 있고, 그 다음으로, "포워드 패스 (forward pass)" 는 출력 (322) 을 생성하기 위하여 연산될 수도 있다. 출력 (322) 은 "표지", "60", 및 "100" 과 같은 특징들에 대응하는 값들의 벡터일 수도 있다. 네트워크 설계자는 DCN 이, 예를 들어, 훈련되었던 네트워크 (300) 에 대한 출력 (322) 에 보여진 바와 같은 "표지" 및 "60" 에 대응하는 것들인, 출력 특징 벡터에서의 뉴런들의 일부에 대한 높은 점수를 출력하기를 원할 수도 있다. 훈련하기 전에, DCN 에 의해 생성된 출력은 올바르지 않을 가능성이 있고, 따라서, 에러가 실제적인 출력과 타겟 출력 사이에서 계산될 수도 있다. 그 다음으로, DCN 의 가중치들은 DCN 의 출력 점수들이 타겟과 더욱 근접하게 정렬되도록 조절될 수도 있다.

[0049] 가중치들을 조절하기 위하여, 학습 알고리즘은 가중치들에 대한 기울기 벡터 (gradient vector) 를 연산할 수도 있다. 기울기는 가중치가 약간 조절되었을 경우에 에러가 증가하거나 감소하는 양을 표시할 수도 있다. 상부 계층에서, 기울기는 마지막에서 두 번째 계층에 있는 활성화된 뉴런 및 출력 계층에 있는 뉴런을 접속하는 가중치의 값에 직접적으로 대응할 수도 있다. 더 낮은 계층들에서, 기울기는 가중치들의 값과, 더 높은 계층들의 연산된 에러 기울기들에 의존할 수도 있다. 그 다음으로, 가중치들은 에러를 감소시키도록 조절될 수도 있다. 가중치들을 조절하는 이 방식은 그것이 신경 네트워크를 통한 "백워드 패스 (backward pass)" 를 수반하므로 "역 전파" 로서 지칭될 수도 있다.

[0050] 실제적으로, 가중치들의 에러 기울기는 계산된 기울기가 참 에러 기울기에 근사하도록, 작은 수의 예들에 대하여 계산될 수도 있다. 이 근사 방법은 확률론적 기울기 하강 (stochastic gradient descent) 으로서 지칭될 수도 있다. 확률론적 기울기 하강은 전체 시스템의 달성가능한 에러 레이트가 감소하는 것이 정지하였을 때까지, 또는 에러 레이트가 타겟 레벨에 도달하였을 때까지 반복될 수도 있다.

[0051] 학습한 후에, DCN 은 새로운 이미지들 (326) 을 제시받을 수도 있고, 네트워크를 통한 포워드 패스는 DCN 의 추론 또는 예측으로 고려될 수도 있는 출력 (322) 을 산출할 수도 있다.

[0052] 심층 신뢰 네트워크 (deep belief network; DBN) 들은 은닉된 노드들의 다수의 계층들을 포함하는 확률적 모델 (probabilistic model) 들이다. DBN 들은 훈련 데이터 세트들의 계층적 표현을 추출하기 위하여 이용될 수도 있다. DBN 은 제한된 볼쯔만 머신 (Restricted Boltzmann Machine; RBM) 들의 계층들을 적층함으로써 획득될 수도 있다. RBM 은 입력들의 세트에 대한 확률 분포를 학습할 수 있는 인공 신경 네트워크의 타입이다. RBM 들은 각각의 입력이 범주화되어야 하는 클래스에 대한 정보의 분재 시에 확률 분포를 학습할 수 있으므로, RBM 들은 비감독된 학습에서 종종 이용된다. 하이브리드의 비감독된 및 감독된 패러다임을 이용하면, DBN 의 하부 RBM 들은 비감독된 방식으로 훈련될 수도 있고 특징 추출기의 역할을 할 수도 있고, 상부 RBM 은 (이전의 계층 및 타겟 클래스들로부터의 입력들의 공동 분포 상에서) 감독된 방식으로 훈련될 수도 있고 분류기로서 작용할 수도 있다.

[0053] 심층 컨볼루션 네트워크 (DCN) 들은 추가적인 풀링 (pooling) 및 정규화 (normalization) 계층들로 구성된 컨볼루션 네트워크들의 네트워크들이다. DCN 들은 다수의 태스크들에 대한 최신 기술의 성능을 달성하였다. DCN 들은 입력 및 출력 타겟 양자 모두가 다수의 견본 (exemplar) 들에 대하여 알려져 있고 기울기 하강 방법들의 이용에 의해 네트워크의 가중치들을 수정하는데 이용되는 감독된 학습을 이용하여 훈련될 수 있다.

[0054] DCN 들은 피드-포워드 네트워크들일 수도 있다. 게다가, 위에서 설명된 바와 같이, DCN 의 제 1 계층에 있는 뉴런으로부터 다음의 더 높은 계층에 있는 뉴런들의 그룹으로의 접속들은 제 1 계층에 있는 뉴런들에 걸쳐 공유된다. DCN 들의 피드-포워드 및 공유된 접속들은 고속 프로세싱을 위하여 활용될 수도 있다. DCN 의 연산 부담은 회귀적 또는 피드백 접속들을 포함하는 유사한 크기의 신경 네트워크의 그것보다 예를 들어, 훨씬 더 작을 수도 있다.

[0055] 컨볼루션 네트워크의 각각의 계층의 프로세싱은 공간적으로 불변인 템플릿 또는 기초 투영 (basis projection) 으로 고려될 수도 있다. 입력이 컬러 이미지의 적색, 녹색, 및 청색 채널들과 같은 다수의 채널들로 먼저 분해될 경우, 그 입력에 대해 훈련된 컨볼루션 네트워크는 이미지의 축들에 따른 2 개의 공간적 차원들 및 컬러 정보를 캡처하는 제 3 차원을 갖는 3 차원으로 고려될 수도 있다. 컨볼루션 접속들의 출력들은 후속 계층 (318 및 320) 에서 특징 맵을 형성하는 것으로 고려될 수도 있고, 특징 맵 (예컨대, 320) 의 각각의 엘리먼트는 이전의 계층 (예컨대, 318) 에 있는 뉴런들의 범위로부터, 그리고 다수의 채널들의 각각으로부터 입력을 수신한다. 특징 맵에서의 값들은 교정 (rectification) 과 같은 비-선형성, max(0,x) 으로 추가로 프로세싱될 수도 있다. 인접한 뉴런들로부터의 값들은 추가로 풀링 (pooling) 될 수도 있는데, 이것은 다운 샘플링에 대응하고, 추가적인 로컬 불변성 및 차원성 감소를 제공할 수도 있다. 백색화 (whitening) 에 대응하는 정규화는 또한, 특징 맵에서의 뉴런들 사이의 측방향 억제 (lateral inhibition) 를 통해 적용될 수도 있다.

[0056] 심층 학습 아키텍처들의 성능은 더 많은 라벨링된 데이터 포인트들이 이용가능해짐에 따라, 또는 연산력이 증가함에 따라 증가할 수도 있다. 최신 심층 신경 네트워크들은 단지 15 년 전에 전형적인 연구자에 의해 이용가능하였던 것보다 수천 배 더 큰 컴퓨팅 자원들로 일상적으로 훈련된다. 새로운 아키텍처들 및 훈련 패러다임들은 심층 학습의 성능을 추가로 상승시킬 수도 있다. 교정된 선형 유닛들은 소실 기울기 (vanishing gradient) 들로서 알려진 훈련 쟁점을 감소시킬 수도 있다. 새로운 훈련 기법들은 오버-피팅 (over-fitting) 을 감소시킬 수도 있고, 이에 따라, 더 큰 모델들이 더욱 양호한 일반화를 달성하는 것을 가능하게 할 수도 있다. 캡슐화 (encapsulation) 기법들은 주어진 수용 필드에서 데이터를 추상화 (abstract) 할 수도 있고, 전체적인 성능을 추가로 상승시킬 수도 있다.

[0057] 도 3b 는 예시적인 심층 컨볼루션 네트워크 (350) 를 예시하는 블록도이다. 심층 컨볼루션 네트워크 (350) 는 접속성 및 가중치 공유에 기초한 다수의 상이한 타입들의 계층들을 포함할 수도 있다. 도 3b 에서 도시된 바와 같이, 예시적인 심층 컨볼루션 네트워크 (350) 는 다수의 컨볼루션 블록들 (예컨대, C1 및 C2) 을 포함한다. 컨볼루션 블록들의 각각은 컨볼루션 계층, 정규화 계층 (LNorm), 및 풀링 계층으로 구성될 수도 있다. 컨볼루션 계층들은 특징 맵을 생성하기 위하여 입력 데이터에 적용될 수도 있는 하나 이상의 컨볼루션 필터들을 포함할 수도 있다. 오직 2 개의 컨볼루션 블록들이 도시되어 있지만, 본 개시물은 그렇게 제한

하지 않고, 그 대신에, 임의의 수의 컨볼루션 블록들이 설계 선호도에 따라 심층 컨볼루션 네트워크 (350) 내에 포함될 수도 있다. 정규화 계층은 컨볼루션 필터들의 출력을 정규화하기 위하여 이용될 수도 있다. 예를 들어, 정규화 계층은 백색화 또는 측방향 억제를 제공할 수도 있다. 풀링 계층은 국소적 불변성 및 차원성 감소를 위하여 공간에 대한 다운 샘플링 어그리게이션 (down sampling aggregation) 을 제공할 수도 있다.

[0058] 심층 컨볼루션 네트워크의 예를 들어, 병렬 필터 뱅크들은 높은 성능 및 낮은 전력 소비를 달성하기 위하여, 임의적으로 ARM 명령 세트에 기초하여, SOC (100) 의 CPU (102) 또는 GPU (104) 상에 로딩될 수도 있다. 대안적인 실시형태들에서, 병렬 필터 뱅크들은 SOC (100) 의 DSP (106) 또는 ISP (116) 상에 로딩될 수도 있다. 게다가, DCN 은 센서들 (114) 및 내비게이션 (120) 에 전용된 프로세싱 블록들과 같은, SOC 상에 존재할 수도 있는 다른 프로세싱 블록들을 액세스할 수도 있다.

[0059] 심층 컨볼루션 네트워크 (350) 는 또한, 하나 이상의 완전히-접속된 계층들 (예컨대, FC1 및 FC2) 을 포함할 수도 있다. 심층 컨볼루션 네트워크 (350) 는 로지스틱 회귀 (logistic regression; LR) 계층을 더 포함할 수도 있다. 심층 컨볼루션 네트워크 (350) 의 각각의 계층 사이에는, 업데이트되어야 하는 가중치들 (도시되지 않음) 이 있다. 각각의 계층의 출력은 제 1 컨볼루션 블록 (C1) 에서 공급된 입력 데이터 (예컨대, 이미지들, 오디오, 비디오, 센서 데이터, 및/또는 다른 입력 데이터) 로부터의 계층적 특징 표현들을 학습하기 위하여 심층 컨볼루션 네트워크 (350) 에서의 연속 계층의 입력의 역할을 할 수도 있다.

[0060] 도 4a 는 인공 지능 (AI) 기능들을 모듈화할 수도 있는 예시적인 소프트웨어 아키텍처 (400) 를 예시하는 블록도이다. 아키텍처를 이용하여, SOC (420) 의 다양한 프로세싱 블록들 (예를 들어, CPU (422), DSP (424), GPU (426), 및/또는 NPU (428)) 로 하여금, 애플리케이션 (402) 의 실행 시간 동작 동안의 연산들을 지원하는 것을 수행하게 할 수도 있는 애플리케이션들 (402) 이 설계될 수도 있다.

[0061] AI 애플리케이션 (402) 은 예를 들어, 디바이스가 현재 동작하는 로케이션을 표시하는 장면의 검출 및 인식을 제공할 수도 있는 사용자 공간 (404) 에서 정의된 함수들을 호출하도록 구성될 수도 있다. AI 애플리케이션 (402) 은 예를 들어, 인식된 장면이 사무실, 강당, 식당, 또는 호수와 같은 실외 세팅인지 여부에 따라 상이하게, 마이크로폰 및 카메라를 구성할 수도 있다. AI 애플리케이션 (402) 은 현재의 장면의 추정을 제공하기 위하여 SceneDetect 애플리케이션 프로그래밍 인터페이스 (application programming interface; API) (406) 에서 정의된 라이브러리와 연관된 컴파일링된 프로그램 코드에 대한 요청을 행할 수도 있다. 이 요청은 예를 들어, 비디오 및 위치결정 데이터에 기초하여 장면 추정치들을 제공하도록 구성된 심층 신경 네트워크의 출력에 궁극적으로 의존할 수도 있다.

[0062] 실행 시간 프레임워크 (Runtime Framework) 의 컴파일링된 코드일 수도 있는 실행 시간 엔진 (408) 은 AI 애플리케이션 (402) 에 의해 추가로 액세스가능할 수도 있다. AI 애플리케이션 (402) 은 실행 시간 엔진으로 하여금, 예를 들어, 특정한 시간 간격으로, 또는 애플리케이션의 사용자 인터페이스에 의해 검출된 이벤트에 의해 트리거링된 장면 추정을 요청하게 할 수도 있다. 장면을 추정하게 될 때, 실행 시간 엔진은 이어서 신호를, SOC (420) 상에서 실행되는 리눅스 커널 (Linux Kernel) (412) 과 같은 오퍼레이팅 시스템 (410) 으로 전송할 수도 있다. 오퍼레이팅 시스템 (410) 은 이어서, 연산이 CPU (422), DSP (424), GPU (426), NPU (428), 또는 그 일부 조합 상에서 수행되게 할 수도 있다. CPU (422) 는 오퍼레이팅 시스템에 의해 직접적으로 액세스될 수도 있고, 다른 프로세싱 블록들은 DSP (424), GPU (426), 또는 NPU (428) 를 위한 드라이버 (414 내지 418) 와 같은 드라이버를 통해 액세스될 수도 있다. 예시적인 예에서, 심층 신경 네트워크는 CPU (422) 및 GPU (426) 와 같은, 프로세싱 블록들의 조합 상에서 실행되도록 구성될 수도 있거나, 존재할 경우, NPU (428) 상에서 실행될 수도 있다.

[0063] 도 4b 는 스마트폰 (452) 상에서의 AI 애플리케이션의 실행 시간 동작 (450) 을 예시하는 블록도이다. AI 애플리케이션은 이미지 (456) 의 포맷을 변환하고 그 다음으로, 이미지 (458) 를 크롭핑하고 및/또는 크기조절 (resize) 하도록 (예를 들어, JAVA 프로그래밍 언어를 이용하여) 구성될 수도 있는 프리-프로세스 (pre-process) 모듈 (454) 을 포함할 수도 있다. 그 다음으로, 프리-프로세싱된 이미지는 시각적 입력에 기초하여 장면들을 검출하고 분류하도록 (예를 들어, C 프로그래밍 언어를 이용하여) 구성될 수도 있는 SceneDetect 백엔드 엔진 (Backend Engine) (462) 을 포함하는 분류 애플리케이션 (460) 으로 통신될 수도 있다. SceneDetect 백엔드 엔진 (462) 은 스케일링 (466) 및 크롭핑 (468) 에 의해 이미지를 추가로 프리프로세싱 (464) 하도록 구성될 수도 있다. 예를 들어, 이미지는 결과적인 이미지가 224 픽셀들 x 224 픽셀들이 되도록 스케일링될 수도 있고 크롭핑될 수도 있다. 이 차원들은 신경 네트워크의 입력 차원들로 맵핑할 수도 있다. 신경 네트워크는 SOC (100) 의 다양한 프로세싱 블록들로 하여금, 심층 신경 네트워크로 이미지 픽셀들

을 추가로 프로세싱하게 하기 위하여 심층 신경 네트워크 블록 (470) 에 의해 구성될 수도 있다. 그 다음으로, 심층 신경 네트워크의 결과들은 임계화 (472) 될 수도 있고 분류 애플리케이션 (460) 에 있는 지수 평활화 블록 (474) 을 통과하게 될 수도 있다. 그 다음으로, 평활화된 결과들은 스마트폰 (452) 의 세팅 및/또는 디스플레이의 변경을 야기시킬 수도 있다.

[0064] 하나의 구성에서, 신경 네트워크와 같은 머신 학습 모델은 네트워크에서의 하나 이상의 계층들을 대체하거나, 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입하거나, 및/또는 압축된 네트워크를 미세-튜닝하기 위하여 구성된다. 모델은 대체 수단, 삽입 수단, 및 튜닝 수단을 포함한다. 하나의 양태에서, 대체 수단, 삽입 수단, 및/또는 튜닝 수단은 범용 프로세서 (102), 범용 프로세서 (102) 와 연관된 프로그램 메모리, 메모리 블록 (118), 로컬 프로세싱 유닛들 (202), 및 또는 인용된 기능들을 수행하도록 구성된 라우팅 접속 프로세싱 유닛들 (216) 일 수도 있다. 또 다른 구성에서, 상기 언급된 수단은 상기 언급된 수단에 의해 인용된 기능들을 수행하도록 구성된 임의의 모듈 또는 임의의 장치일 수도 있다.

[0065] 다른 구성에서, 신경 네트워크와 같은 머신 학습 모델은 압축된 네트워크를 생성하기 위하여, 네트워크에서의 하나 이상의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 하기 위하여 구성된다. 모델은 또한, 압축된 계층들 중의 하나 이상에서 가중치들을 업데이트함으로써 압축된 네트워크를 미세-튜닝하기 위하여 구성된다. 모델은 대체 수단 및 튜닝 수단을 포함한다. 하나의 양태에서, 대체 수단 및/또는 튜닝 수단은 범용 프로세서 (102), 범용 프로세서 (102) 와 연관된 프로그램 메모리, 메모리 블록 (118), 로컬 프로세싱 유닛들 (202), 및 또는 인용된 기능들을 수행하도록 구성된 라우팅 접속 프로세싱 유닛들 (216) 일 수도 있다. 다른 구성에서, 상기 언급된 수단은 상기 언급된 수단에 의해 인용된 기능들을 수행하도록 구성된 임의의 모듈 또는 임의의 장치일 수도 있다.

[0066] 또 다른 구성에서, 신경 네트워크와 같은 머신 학습 모델은 네트워크에서의 하나 이상의 계층들을 대체하거나, 및/또는 교차 최소화 프로세스를 적용함으로써 압축된 계층들의 가중치 매트릭스들을 결정하기 위하여 구성된다. 모델은 대체 수단 및 결정 수단을 포함한다. 하나의 양태에서, 대체 수단 및/또는 결정 수단은 범용 프로세서 (102), 범용 프로세서 (102) 와 연관된 프로그램 메모리, 메모리 블록 (118), 로컬 프로세싱 유닛들 (202), 및 또는 인용된 기능들을 수행하도록 구성된 라우팅 접속 프로세싱 유닛들 (216) 일 수도 있다. 또 다른 구성에서, 상기 언급된 수단은 상기 언급된 수단에 의해 인용된 기능들을 수행하도록 구성된 임의의 모듈 또는 임의의 장치일 수도 있다.

[0067] 본 개시물의 어떤 양태들에 따르면, 각각의 로컬 프로세싱 유닛 (202) 은 네트워크의 바람직한 하나 이상의 기능적 특징들에 기초하여 네트워크의 파라미터들을 결정하고, 결정된 파라미터들이 추가로 적응되고, 튜닝되고, 업데이트될 때, 바람직한 기능적 특징들을 향해 하나 이상의 기능적 특징들을 개발하도록 구성될 수도 있다.

[0068] 완전히-접속된 계층들의 압축

[0069] 도 5a 내지 도 5b 는 본 개시물의 양태들에 따라 완전히-접속된 계층 및 압축된 완전히-접속된 계층을 예시하는 블록도들이다. 도 5a 에서 도시된 바와 같이, 완전히-접속된 계층 (fc) (500) 의 출력 뉴런들 (504) 의 각각은 올-투-올 (all-to-all) 방식으로 시냅스 (synapse) 를 통해 입력 뉴런들 (502) 의 각각에 접속된다 (올-투-올 접속들의 서브세트는 예시의 용이함을 위하여 도시됨). fc 계층은 n개 입력 뉴런들 및 m개 출력 뉴런들을 가진다. i 번째 입력 뉴런을 j 번째 출력 뉴런에 접속하는 시냅스의 가중치는 W_{ij} 에 의해 나타내어진다.

n개 입력 뉴런들 및 m개 출력 뉴런들을 접속하는 시냅스들의 가중치들의 전부는 매트릭스 $\mathbb{W}$ 를 이용하여 집합적으로 표현될 수도 있다. 출력 활성화 벡터 y 는 입력 활성화 벡터 x 를 가중치 매트릭스 $\mathbb{W}$ 에 의해 승산함으로써, 그리고 바이어스 벡터 b 를 가산함으로써 획득될 수도 있다.

[0070] 도 5b 는 예시적인 압축된 완전히-접속된 계층 (550) 을 예시한다. 도 5b 에서 도시된 바와 같이, 단일의 완전히-접속된 계층은 2 개의 완전히-접속된 계층들로 대체된다. 즉, 도 5a 의 'fc' 로 명명된 완전히-접속된 계층은 도 5b 에서 도시된 2 개의 완전히-접속된 계층들 'fc1' 및 'fc2' 로 대체된다. 오직 하나의 추가적인 완전히-접속된 계층이 도시되어 있지만, 본 개시물은 그렇게 제한되지 않고, 추가적인 완전히-접속된 계층들이 완전히-접속된 계층 fc 를 압축하기 위하여 이용될 수도 있다. 예를 들어, 도 6a 내지 도 6b 는 완전히-접속된 계층, 및 3 개의 완전히-접속된 계층들을 갖는 압축된 완전히-접속된 계층을 예시한다. 도 6a 내지 도 6b 에서 도시된 바와 같이, 비압축된 완전히-접속된 계층 'fc' 는 3 개의 완전히-접속된 계층들 'fc1', 'fc2', 및 'fc3' 으로 대체된다. 중간 계층들에서의 뉴런들 r_1 및 r_2 의 수는 그것들이 달성하는 압축과, 결과적인 압축된 네트워크의 성능에 기초하여 결정될 수도 있다.

[0071] 압축에 있어서의 장점은 압축 전 및 후의 파라미터들의 총 수를 비교함으로써 이해될 수 있다. 압축 전의 파라미터들의 수는 가중치 매트릭스 W = nm 에서의 엘리먼트들의 수 플러스 (plus) m 과 동일한 바이어스 벡터와 동일할 수도 있다. 이에 따라, 압축 전의 파라미터들의 총 수는 nm + m 과 동일하다. 그러나, 도 5b 에서의 압축 후의 파라미터들의 수는 nr + rm 과 동일한, 계층들 'fc1' 및 'fc2' 에서의 파라미터들의 수의 합과 동일하다. r 의 값에 따라, 파라미터들의 수에 있어서의 상당한 감소가 달성될 수 있다.

[0072] 압축된 네트워크에 의해 달성된 유효 변환은 이하에 의해 주어진다:

[0073] $$y = W_2 W_1 x + b \quad (1)$$

[0074] 식중, W_1 및 W_2 는 각각 압축된 완전히-접속된 계층들 fc1 및 fc2 에 대한 가중치 매트릭스들이다.

[0075] 압축된 계층들의 가중치 매트릭스들은 유효 변환이 원래의 완전히-접속된 계층 fc 에 의해 달성된 변환의 근접한 근사화가 되도록 결정될 수도 있다:

[0076] $$y = Wx + b \quad (2)$$

[0077] 가중치 매트릭스들 W_1 및 W_2 는 이하에 의해 주어지는 근사화 에러를 감소시키도록 선택될 수도 있어서:

[0078] $$|W - W_2 W_1|^2 \quad (3)$$

[0079] 압축된 계층들 fc1 및 fc2 에 의해 달성된 수학식 1 의 유효 변환은 수학식 2 에 명시된 원래의 완전히-접속된 계층의 변환을 근접하게 근사화한다.

[0080] 유사하게, 도 6b 에서의 압축 후의 파라미터들의 수는 $nr_1 + r_1 r_2 + r_2 m + m$ 과 동일하다. 또한, r_1 및 r_2 의 값들에 따라, 파라미터들의 수에 있어서의 상당한 감소가 달성될 수 있다.

[0081] 도 6b 에서 도시된 압축된 네트워크에서의 압축된 계층들 'fc1', 'fc2', 및 'fc3' 에 의해 함께 달성된 유효 변환은 $y = W_3 W_2 W_1 x + b$ 이다. 가중치 매트릭스들 W_1, W_2, 및 W_3 은 유효 변환이 (도 6a 에서 도시된) 원래의 'fc' 계층에 의해 달성된 변환의 근사화 $y = Wx + b$ 가 되도록 결정될 수도 있다.

[0082] 압축된 계층들의 가중치 매트릭스들을 결정하기 위한 하나의 방법은 특이 값 분해 (singular value decomposition; SVD) 를 반복적으로 적용하는 것에 의한 것이다. SVD 는 가중치 매트릭스들 W_1 및 W_2 이 결정될 수도 있는 W 의 랭크 근사 (rank approximation) 를 연산하기 위하여 이용될 수도 있다.

[0083] 압축된 계층들의 가중치 매트릭스들 W_1 및 W_2 을 결정하기 위한 또 다른 방법은 교차 최소화 프로세스를 적용하는 것에 의한 것이다. 교차 최소화 프로세스에 따르면, 매트릭스들은 랜덤 값들로 초기화되고, 압축된 매트릭스들은 수학식 3 의 목적 함수를 감소시키거나 최소화함으로써 근사화를 개선시키기 위하여 최소 제곱 (least square) 들을 이용하여 한 번에 하나씩 교번하여 업데이트된다.

[0084] 일부 양태들에서, 가중치 매트릭스들은 압축된 네트워크를 훈련함으로써 결정될 수도 있다. 예를 들어, 비압축된 계층에 대한 입력들 및 출력들은 훈련 예들의 세트에 대하여 기록될 수도 있다. 그 다음으로, 예를 들어, 경사 하강은 압축된 계층들을 초기화하여 비압축된 계층에 대응하는 출력들을 생성하기 위하여 이용될 수도 있다.

[0085] 압축된 네트워크는 원래의 네트워크의 근사이므로, 압축된 네트워크의 종단-대-종단 (end-to-end) 거동은 원래의 네트워크에 비해 약간 또는 상당히 상이할 수도 있다. 그 결과, 네트워크가 달성하도록 설계되는 태스크에 대하여, 원래의 네트워크뿐만 아니라 새로운 압축된 네트워크도 수행하지 않을 수도 있다. 성능에 있어서의 하락을 방지하기 위하여, 압축된 네트워크의 가중치들을 미세-튜닝하고 수정하는 것이 수행될 수도 있다.

예를 들어, 압축된 계층들의 전부를 통한 역 전파는 가중치들을 수정하기 위하여 적용될 수도 있다. 대안적으로, 역 전파는 계층들의 서브세트를 통해 수행될 수도 있거나, 출력으로부터 다시 입력으로의 전체 모델의 전반에 걸친 압축 및 비압축된 계층들의 양자를 통해 수행될 수도 있다. 미세-튜닝은 일부 훈련 예들로 신경 네트워크를 교육함으로써 달성될 수 있다. 이용가능성에 따라서는, 원래의 네트워크를 훈련하기 위한 동일한 훈련 예들, 또는 훈련 예들의 상이한 세트, 또는 구 및 신 훈련 예들의 서브세트 중의 어느 하나가 채용될 수도 있다.

[0086] 미세-튜닝 동안, 우리는 오직 새로운 압축된 계층들, 또는 (전체 스택 미세-튜닝으로서 지칭될 수도 있는) 모든 계층들을 함께, 또는 압축 및 비압축된 계층들의 서브세트를 미세-튜닝하는 것을 선택할 수도 있다. 예를 들어, 도 5b 및 도 6b 에서 도시된 바와 같은 압축된 계층들은 도 5a 및 도 6a 에서 각각 도시된 바와 같은 완전히-접속된 계층 타입의 또 다른 사례이다. 이에 따라, 압축된 계층들을 훈련하거나, 또는 압축된 네트워크를 이용하여 추론을 수행하기 위한 것 중의 어느 하나를 위한 새로운 기법들의 설계 또는 구현은 회피될 수도 있다. 그 대신에, 원래의 네트워크를 위하여 이용가능한 임의의 훈련 및 추론 플랫폼이 재이용될 수도 있다.

[0087] 컨볼루션 계층들의 압축

[0088] 신경 네트워크의 컨볼루션 계층들이 또한 압축될 수도 있다. 컨볼루션 계층들을 압축하기 위한 접근법은 여러 가지 점에서, 완전히-접속된 계층들에 대하여 위에서 설명된 것과 유사하다. 예를 들어, 완전히-접속된 계층들의 경우에서와 같이, 압축된 컨볼루션 계층들의 가중치 매트릭스들은 압축된 계층들에 의해 달성된 유효 변환이 원래의 비압축된 컨볼루션 계층에 의해 달성된 변환에 대한 양호한 근사화가 되도록 선택될 수도 있다. 그러나, 완전히-접속된 계층들과 컨볼루션 계층들 사이의 아키텍처 차이들로 인한 일부 차이들이 있다.

[0089] 도 7 은 본 개시물의 양태들에 따라 예시적인 컨볼루션 계층을 예시하는 블록도이다. 컨볼루션 계층은 n개 입력 맵들을 m개 출력 맵들에 접속시킨다. 각각의 맵은 이미지의 공간적 차원들, 또는 오디오 신호에서의 시간-주파수 차원들 등 중의 어느 하나에 대응하는 뉴런들의 세트 (예컨대, 2-차원 (2-D) 그리드 (grid) 또는 3-D 맵) 를 포함할 수도 있다. 예를 들어, 일부 양태들에서, 뉴런들의 세트는 비디오 데이터 또는 복셀 데이터 (voxel data) (예컨대, 자기 공명 이미징 (magnetic resonance imaging; MRI) 스캔) 에 대응할 수도 있다. 컨볼루션 계층은 컨볼루션 커널 W^{ij} 를 i 번째 입력 맵 X^i 에 적용하고, 그 결과를 j 번째 출력 맵 Y^j 에 추가하고, 여기서, 인덱스 i 는 1 로부터 n 까지 진행된다. 다시 말해서, j 번째 출력 맵 Y^j 은 이하로서 표현될 수 있다:

[0090] $$Y^j = \sum_{i=1}^{n} X^i * W^{ij} + B^j \quad (4)$$

[0091] 도 8a 내지 도 8b 는 본 개시물의 양태들에 따라 컨볼루션 계층의 일 예의 압축을 예시한다. 도 8a 내지 도 8b 를 참조하면, 'conv' 로 명명된 컨볼루션 계층은 2 개의 컨볼루션 계층들 'conv1' 및 'conv2' 로 대체된다.

[0092] 완전히-접속된 계층들에 대하여 위에서 표시된 바와 같이, 압축된 계층들 (대체 계층들) 의 수는 도 8b 에서 도시된 2 로 제한되지 않고, 임의의 수의 압축된 계층들이 설계 선호도에 따라 그 대신에 이용될 수도 있다. 예를 들어, 도 9a 내지 도 9b 는 'conv' 계층이 3 개의 컨볼루션 계층들 'conv1', 'conv2', 및 'conv3' 로 대체되는 일 예의 압축을 예시한다. 그 경우 (도 9b), 중간 계층들에서의 맵들 r_1 및 r_2 의 수는 달성된 압축, 및 결과적인 압축된 네트워크의 성능에 기초하여 결정될 수도 있다.

[0093] 2 개의 컨볼루션 계층들로 대체

[0094] 도 8b 에서 설명된 압축된 네트워크에서 압축된 계층들 'conv1' 및 'conv2' 에 의해 함께 달성되는 유효 변환은 이하로서 표현될 수 있다:

[0095] $$Y^j = \sum_{k=1}^{r_1} Z^k * W_2^{kj} + B^j = \sum_{i=1}^{n} X^i * \left(\sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj} \right) + B^j \quad (5)$$

[0096] 식중, 이 예에서 Bj 는 바이어스 항이고, Z^k 는 뉴런들의 추가된 계층의 활성화 벡터이고, $r_1 = r$ 이다. 본질적으로, 원래의 네트워크에서의 'conv' 계층의 가중치 매트릭스들 W^{ij} 는 'conv1' 계층의 가중치 매트릭스들 W_1^{ik} 및 'conv2' 계층의 W_2^{kj} 을 통해 압축된 네트워크에서 이하로서 근사화된다

[0097] $$W^{ij} \approx \sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}, \; i = 1,2 \ldots, n, \; j = 1,2, \ldots, m. \quad (6)$$

[0098] 확실히, 압축된 컨볼루션 계층들의 유효 변환 (수학식 5) 이 원래의 컨볼루션 계층의 변환에 대한 근접한 근사이도록 하기 위하여, 압축된 계층들의 가중치 매트릭스들은 이하로서 명시되는 근사화 에러를 감소시키거나 최소화하도록 선택될 수도 있다:

[0099] $$\sum_{i,j=1,1}^{n,m}\left(W^{ij}-\sum_{k=1}^{r_1}W_1^{ik}*W_2^{kj}\right)^2 \quad . \qquad (7)$$

[0100] 또한, SVD 방법, 교차 최소화 프로세스 등등은 수학식 7 의 목적 함수를 감소시키거나 최소화하는 압축된 계층의 가중치 매트릭스들을 결정하기 위하여 이용될 수도 있다.

[0101] 압축된 컨볼루션 계층들의 커널 크기들

[0102] 수학식 7 에서 언급된 근사화 에러가 타당하기 위하여, 매트릭스 W_{ij} 는 매트릭스 $W_1^{ik}*W_2^{kj}$ 와 동일한 차원들을 가져야 한다. 이 목적을 위하여, 압축된 컨볼루션 계층들 'conv1' 및 'conv2' 의 커널 크기들이 적절하게 선택될 수도 있다. 예를 들어, $k_x \times k_y$ 가 비압축된 'conv' 계층의 커널 크기를 나타낼 경우, 압축된 컨볼루션 계층들 'conv1' 및 'conv2' 의 커널 크기는 그것들이 이하를 충족시키도록 선택될 수도 있다:

[0103] $$(k_{1x}-1)+(k_{2x}-1)=(k_x-1)$$

[0104] 및 (8)

[0105] $$(k_{1y}-1)+(k_{2y}-1)=(k_y-1),$$

[0106] 식중, $k_{1x} \times k_{1y}$ 은 'conv1' 계층의 커널 크기를 나타내고, $k_{2x} \times k_{2y}$ 은 'conv2' 계층의 커널 크기를 나타낸다.

[0107] 이에 따라, 커널 크기 $k_x \times k_y$ 를 갖는 컨볼루션 계층은 각각 커널 크기들 $k_x \times k_y$ 및 1 x 1 을 갖는 2 개의 컨볼루션 계층들로 대체될 수도 있고, 그 역 또한 마찬가지이다.

[0108] 또한, 일부 양태들에서, 커널 크기 $k_x \times k_y$ 를 갖는 컨볼루션 계층은 커널 크기들 1 x 1, $k_x \times k_y$, 및 1 x 1 을 각각 갖는 3 개의 컨볼루션 계층들로 대체될 수도 있다.

[0109] 일부 양태들에서, 압축된 컨볼루션 계층들, 'conv1' 및 'conv2' 의 개개의 커널 크기들 k_1 및 k_2 는 이하를 충족시키도록 선택될 수도 있다:

[0110] $$k_1+k_2-1=k, \qquad (9)$$

[0111] 식중, k 는 비압축된 컨볼루션 계층 'conv' 의 커널 크기이다. 이하의 표 1 은 상이한 가능한 커널 크기들 k_1 및 k_2 의 예를 제공하고, 여기서, 비압축된 컨볼루션 계층 'conv' 의 커널 크기는 k=5 이다.

표 1

[0112]

k_1	k_2	k_1+k_2-1
5	1	5
4	2	5
3	3	5
2	4	5
1	5	5

[0113] 다수의 컨볼루션 계층들로 대체

[0114] 컨볼루션 계층은 또한, 2 개를 초과하는 컨볼루션 계층들로 대체될 수 있다. 설명의 용이함을 위하여, L 컨볼루션 계층들이 하나의 컨볼루션 계층을 대체하기 위하여 이용되는 것으로 가정한다. 그 다음으로, l 번째

계층의 가중치 값들 W_l 은 목적 함수를 감소시키거나 최소화하기 위하여 선택되고,

[0115] $$\sum_{i,j=1,1}^{n,m}\left(W^{ij}-\sum_{k_1,k_2\ldots k_L=1,1,\ldots 1}^{r_1,r_2,\cdots,r_L} W_1^{ik_1} * W_2^{k_1k_2} * \cdots * W_L^{k_{L-1},j}\right)^2 , \quad (10)$$

[0116] 식중, 변수들 r_l 은 l 번째 계층의 출력 맵들의 수를 나타낸다. 또한, 목적 함수가 타당하기 위하여, l 번째 계층의 커널 크기 $k_{lx} \times k_{ly}$ 는 그것들이 이하를 충족시키도록 선택될 수도 있다:

[0117] $$(k_{1x}-1)+(k_{2x}-1)+\cdots+(k_{Lx}-1)=(k_x-1)$$

[0118] 및 (11)

[0119] $$(k_{1y}-1)+(k_{2y}-1)+\cdots+(K_{Lx}-1)=(k_y-1).$$

[0120] 따라서, 일부 양태들에서, 커널 크기 $k_x \times k_y$ 를 갖는 컨볼루션 계층은, 특성들 $(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$ 및 $(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$ 이 충족되도록, 커널 크기들 $k_{1x} \times k_{1y}$, $k_{2x} \times k_{2y}$ 등등을 갖는 다수의 컨볼루션 계층들로 대체된다.

[0121] 교차 최소화 프로세스

[0122] 압축된 컨볼루션 계층들 중의 하나가 1 x 1 의 커널 크기를 가지고, 다른 압축된 계층이 비압축된 계층의 동일한 커널 크기를 가지는 특수한 경우에 있어서, SVD 방법은 예를 들어, 수학식 7 및 수학식 10 에서 기재된 목적 함수를 최적화하는 압축된 가중치 매트릭스들을 결정하기 위하여 이용될 수 있다. 하나의 양태에서, 이 특수한 경우는 ($k_{1x} \times k_{1y}=1\times 1$ 및 $k_{2x} \times k_{2y}=k_x \times k_y$) 또는 ($k_{1x} \times k_{1y}=k_x \times k_y$ 및 $k_{2x} \times k_{2y}=1\times 1$) 의 어느 하나에 대응한다.

[0123] 압축된 계층이 1 x 1 의 커널 크기를 가지지 않는 일반적인 경우에 있어서, 교차 최소화 방법은 수학식 7 및 수학식 10 을 최적화하는 압축된 가중치 매트릭스들을 결정하기 위하여 이용될 수 있다. 교차 최소화 방법은 다음의 단계들을 포함한다:

[0124] (1) 랜덤 값들로 'conv1' 계층의 가중치 매트릭스들 W_1^{ik} 및 'conv2' 계층의 W_2^{kj} 을 초기화

[0125] (2) 교번 방식으로 각각의 압축된 계층의 가중치들에 대해 푼다:

[0126] (a) 'conv2' 계층 가중치들을 고정하고, 수학식 7 을 최소화하는 'conv1' 계층 가중치들에 대해 푼다.

[0127] (b) 'conv1' 계층 가중치들을 고정하고, 수학식 7 을 최소화하는 'conv2' 계층 가중치들에 대해 푼다.

[0128] (3) 수렴시까지, 또는 미리 결정된 수의 반복들에 대하여 단계 2 를 반복함.

[0129] 단계들 2a 및 2b 는 표준 최소-제곱 프로세스를 이용하여 닫힌-형태로 풀어질 수 있다.

[0130] 압축된 가중치들이 SVD 방법 또는 교차 최소화 방법을 이용하여 결정된 후, 압축된 네트워크는 분류 성능에 있어서의 손실을 어느 한도까지 되찾기 위하여 미세-튜닝될 수 있다는 점에 유의한다.

[0131] 국소적으로-접속된 계층들의 압축

[0132] 컨볼루션 계층에서, 동일한 컨볼루션 커널은 입력 이미지 맵의 전반에 걸쳐 적용되는 반면, 국소적으로-접속된 계층에서는, 상이한 가중치들이 상이한 공간적 로케이션들에서 적용된다. 따라서, 컨볼루션 계층들에 대하여 설명된 압축 접근법은 상이한 공간적 로케이션들에서 압축 방법 (상기 수학식들 4 내지 11 을 참조) 을 적용함으로써 국소적으로-접속된 계층들에 유사하게 적용될 수 있다.

[0133] 설계 파라미터들의 선택

[0134] 압축에 수반된 여러 설계 파라미터들 (예컨대, 완전히-접속된 계층들의 경우에 일정 수의 중간 뉴런들; 컨볼루션 계층들의 경우에 일정 수의 출력 맵들 및 커널 크기) 은 예를 들어, 달성된 압축 및 압축된 네트워크의 대응하는 기능적 성능을 경험적으로 측정함으로써 선택될 수도 있다.

[0135] 하나의 예에서, 중간 계층에 있는 뉴런들의 수 r 은 파라미터 스윕 (parameter sweep) 을 이용하여 선택될 수도 있다. r 의 값은 16 으로부터 min(n, m) 까지 16 을 증분 (increment) 으로 스윕 (sweep) 될 수도 있다. 각각의 값에 대하여, 유효성 데이터 세트에 대한 압축된 네트워크의 분류 정확도가 결정될 수도 있다. 분류 정확도에 있어서의 하락이 수용가능한 (예컨대, 임계치 미만) r 의 최저 값이 선택될 수도 있다.

[0136] 일부 양태들에서, "그리디 탐색 (greedy search) ' 방법이 설계 파라미터들을 결정하기 위하여 이용될 수도 있다. 예를 들어, 압축을 위한 가장 많은 가능성을 보이는 계층들 (예컨대, 많은 수의 연산들을 갖는 계층들) 중의 하나 이상의 압축될 수도 있다. 그 후에, 압축을 위한 추가적인 계층들은 선택적으로 압축될 수도 있다.

[0137] 일부 양태들에서, 양호한 설계 파라미터들은 각각의 계층을 개별적으로 압축함으로써 결정될 수도 있다. 예를 들어, 각각의 계층은 기능적 성능이 임계치 미만으로 하락하지 않도록, 가능한 최대 한도로 압축될 수도 있다. 그 다음으로, 각각의 계층에 대하여 개별적으로 학습된 이 설계 파라미터들은 네트워크에서의 다수의 계층들을 함께 압축하기 위하여 이용될 수도 있다. 다수의 계층들을 함께 압축한 후, 네트워크는 기능적 성능에 있어서의 하락을 되찾기 위하여 압축 및 비압축된 계층들에서 가중치들을 업데이트함으로써 미세-튜닝될 수도 있다.

[0138] 일부 양태들에서, 압축된 네트워크는 각각의 결정된 파라미터 값 (예컨대, *r*) 에 대하여 미세-튜닝될 수도 있거나, 오직 최종 선택된 파라미터 값에 대하여 미세-튜닝될 수도 있다. 다음의 의사-코드는 압축 파라미터들을 선택하기 위한 예시적인 휴리스틱 (heuristic) 을 제공한다.

[0139] 각각의 완전히-접속된 계층에 대하여:

[0140] max_rank = min(n,m)

[0141] r = 1:max_rank 에 대하여

[0142] 선택된 완전히-접속된 계층을 압축 (다른 것들은 비압축된 상태로 남김)

[0143] 분류 정확도를 결정

[0144] 정확도에 있어서의 하락이 임계치 미만이 되도록 최저 r 을 선택

[0145] 대응하는 선택된 r 값들을 이용하여 모든 완전히-접속된 계층들을 압축

[0146] 압축된 네트워크를 미세-튜닝.

[0147] 압축된 계층들 사이의 비선형 계층들의 삽입

[0148] 계층을 다수의 계층들로 대체한 후, 압축된 계층에 있는 뉴런들은 아이덴티티 활성화들로 구성될 수도 있다. 그러나, 비선형 계층들은 네트워크의 표현 용량을 개선시키기 위하여 압축된 계층들 사이에 추가될 수도 있다. 또한, 비선형 계층들을 추가함으로써, 더 높은 기능적 성능이 달성될 수도 있다.

[0149] 하나의 예시적인 양태에서는, 교정기 비선형성이 삽입될 수도 있다. 이 예에서는, 그것들을 다음 계층으로 전달하기 전에, 임계치가 max(0,x) 비선형성을 이용하여 압축된 계층의 출력 활성화에 적용될 수도 있다. 일부 양태들에서, 교정기 비선형성은 교정된 선형 유닛 (rectified linear unit; ReLU) 일 수도 있다. abs(x), sigmoid(x), 및 쌍곡선 탄젠트 (hyperbolic tangent) (tanh (x)) 등등을 포함하는 다른 종류들의 비선형성들이 또한 이용될 수도 있다.

[0150] 일부 양태들에서는, 신경 네트워크가 압축될 수도 있거나, 비-선형 계층이 삽입될 수도 있거나, 및/또는 미세-튜닝이 적용될 수도 있다. 예를 들어, 9 x 9 컨볼루션 계층을 갖는 신경 네트워크는 (예컨대, 훈련 예들의 제 1 세트를 이용하여) 훈련될 수도 있다. 그 다음으로, 9 x 9 컨볼루션 계층은 그것을 2 개의 5 x 5 컨볼루션 계층들로 대체하여 압축될 수도 있다. 비선형 계층은 2 개의 5 x 5 컨볼루션 계층들 사이에 추가될 수도 있다. 미세-튜닝이 수행될 수도 있다. 일부 양태들에서, 그 다음으로, 네트워크는 5 x 5 컨볼루션 계층들의 각각을 2 개의 3 x 3 컨볼루션 계층들로 대체함으로써 추가로 압축될 수도 있다. 또한, 비선형 계층은 3 x 3 컨볼루션 계층들의 각각 사이에 추가될 수도 있다.

[0151] 그 다음으로, 추가적인 미세-튜닝 및 압축은 바람직한 성능이 획득될 때까지 반복될 수도 있다. 상기 예에서, 원래의 네트워크와 비교하여, 최종 압축된 네트워크는 하나의 9 x 9 컨볼루션 계층 대신에, 비선형 계층들을 사이에 갖는 4 개의 3 x 3 컨볼루션 계층들을 가진다. 3 개의 추가적인 컨볼루션 및 비선형 계층들을 갖는 더욱 심층 네트워크는 더욱 양호한 기능적 성능을 달성할 수도 있다.

[0152] 도 10 은 신경 네트워크를 압축하기 위한 방법 (1000) 을 예시한다. 블록 (1002) 에서, 프로세스는 압축된 네트워크를 생성하기 위하여 신경 네트워크에서의 하나 이상의 계층들을 다수의 압축된 계층들로 대체한다. 일부 양태들에서, 압축된 계층들은 초기 비압축된 네트워크에서의 계층들과 동일한 타입이다. 예를 들어, 완전히-접속된 계층은 다수의 완전히-접속된 계층들에 의해 대체될 수도 있고, 컨볼루션 계층은 다수의 컨볼루션 계층들에 의해 대체될 수도 있다. 마찬가지로, 국소적으로-접속된 계층은 다수의 국소적으로-접속된 계층들에 의해 대체될 수도 있다.

[0153] 블록 (1004) 에서, 프로세스는 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입한다. 일부 양태들에서, 프로세스는 비선형 활성화 함수를 압축된 계층들의 뉴런들에 적용함으로써 비선형성을 삽입한다. 비선형 활성화 함수는 교정기, 절대 값 함수, 쌍곡선 탄젠트 함수, 시그모이드 함수 (sigmoid function), 또는 다른 비선형 활성화 함수를 포함할 수도 있다.

[0154] 또한, 블록 (1006) 에서, 프로세스는 계층들 중의 적어도 하나에서 가중치들을 업데이트함으로써 압축된 네트워크를 미세-튜닝한다. 미세-튜닝은 역 전파 또는 다른 표준 프로세스들을 이용하여 수행될 수도 있다. 미세-튜닝은 압축된 신경 네트워크에서 모든 가중치들을 업데이트함으로써, 또는 압축된 계층들의 서브세트, 비압축된 계층들의 서브세트, 또는 양자의 혼합을 업데이트함으로써 수행될 수도 있다. 미세-튜닝은 훈련 예들을 이용하여 수행될 수도 있다. 훈련 예들은 원래의 신경 네트워크를 교육하기 위하여 이용된 것들과 동일할 수도 있거나, 예들의 새로운 세트, 또는 양자의 혼합을 포함할 수도 있다.

[0155] 일부 양태들에서, 훈련 예들은 예를 들어, 사진들의 갤러리를 포함하는, 스마트폰 또는 다른 이동 디바이스로부터의 데이터를 포함할 수도 있다.

[0156] 일부 양태들에서, 프로세스는 압축을 반복적으로 적용하는 것, 비선형 계층들의 삽입, 및 더욱 심층 신경 네트워크들을 초기화하기 위한 방법으로서의 미세-튜닝에 의해 신경 네트워크를 초기화하는 것을 더 포함할 수도 있다.

[0157] 도 11 은 신경 네트워크를 압축하기 위한 방법 (1100) 을 예시한다. 블록 (1102) 에서, 프로세스는 압축된 네트워크를 생성하기 위하여, 신경 네트워크에서의 적어도 하나의 계층을 다수의 압축된 계층들로 대체하여 조합된 압축된 계층들의 수용 필드 크기가 비압축된 계층들의 수용 필드와 일치하도록 한다. 일부 양태들에서, 비압축된 계층들의 커널 크기는 수용 필드 크기와 동일하다.

[0158] 또한, 블록 (1104) 에서, 프로세스는 압축된 계층들 중의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝한다.

[0159] 도 12 는 신경 네트워크를 압축하기 위한 방법 (1200) 을 예시한다. 블록 (1202) 에서, 프로세스는 압축된 네트워크를 생성하기 위하여 신경 네트워크에서의 하나 이상의 계층들을 다수의 압축된 계층들로 대체한다. 일부 양태들에서, 압축된 계층들은 초기 비압축된 네트워크에서의 계층들과 동일한 타입이다. 예를 들어, 완전히-접속된 계층은 다수의 완전히-접속된 계층들에 의해 대체될 수도 있고, 컨볼루션 계층은 다수의 컨볼루션 계층들에 의해 대체될 수도 있다. 마찬가지로, 국소적으로-접속된 계층은 다수의 국소적으로-접속된 계층들에 의해 대체될 수도 있다.

[0160] 또한, 블록 (1204) 에서, 프로세스는 교차 최소화 프로세스를 적용함으로써 다수의 압축된 계층들의 가중치 매트릭스들을 결정한다.

[0161] 일부 양태들에서, 프로세스는 또한, 압축된 계층들의 적어도 하나에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝한다. 미세-튜닝은 압축된 신경 네트워크에서 모든 가중치들을 업데이트함으로써, 또는 압축된 계층들의 서브세트, 비압축된 계층들의 서브세트, 또는 양자의 혼합을 업데이트함으로써 수행될 수도 있다. 미세-튜닝은 훈련 예들을 이용하여 수행될 수도 있다. 훈련 예들은 원래의 신경 네트워크를 교육하기 위하여 이용된 것들과 동일할 수도 있거나, 예들의 새로운 세트, 또는 양자의 혼합을 포함할 수도 있다.

[0162] 미세-튜닝은 단일 스테이지에서 수행될 수도 있거나, 다수의 스테이지들에서 수행될 수도 있다. 예를 들어, 미세-튜닝이 다수의 스테이지들에서 수행될 때, 제 1 스테이지에서, 미세-튜닝은 오직 압축된 계층들의 서브세

트에 대하여 수행된다. 그 다음으로, 제 2 스테이지에서, 미세-튜닝은 압축 및 비압축된 계층들의 서브세트에 대하여 수행된다.

[0163] 도 13 은 본 개시물의 양태들에 따라 신경 네트워크를 압축하기 위한 방법 (1300) 을 예시한다. 블록 (1302) 에서, 프로세스는 다수의 차원들을 따르는 신경 네트워크와 같은 머신 학습 네트워크를 특성화한다. 일부 양태들에서는, 신경 네트워크의 태스크 수행 (P) 이 특성화될 수도 있다. 태스크 성능은 신경 네트워크가 객체 분류기일 경우에 테스팅 세트에 대한 올바르게 분류된 객체들의 백분율일 수도 있다. 메모리 풋프린트 (M) 및/또는 추론 시간이 또한 특성화될 수도 있다. 메모리 풋프린트는 예를 들어, 신경 네트워크의 가중치들 및 다른 모델 파라미터들의 메모리 저장 요건들에 대응할 수도 있다. 추론 시간 (T) 은 그것이 신경 네트워크에 제시된 후에 새로운 객체의 분류 동안에 경과된 시간일 수도 있다.

[0164] 블록 (1304) 에서, 프로세스는 압축된 네트워크를 생성하기 위하여 신경 네트워크에서의 하나 이상의 계층들을 다수의 압축된 계층들로 대체한다. 일부 양태들에서, 압축된 계층들은 초기 비압축된 네트워크에서의 계층들과 동일한 타입일 수도 있다. 예를 들어, 완전히-접속된 계층은 다수의 완전히-접속된 계층들에 의해 대체될 수도 있고, 컨볼루션 계층은 다수의 컨볼루션 계층들에 의해 대체될 수도 있다. 마찬가지로, 국소적으로-접속된 계층은 다수의 국소적으로-접속된 계층들에 의해 대체될 수도 있다. 예시적인 구성에서, 프로세스는 신경 네트워크의 제 1 계층을 제 2 및 제 3 계층으로 대체하여, 제 3 계층의 크기가 제 1 계층의 크기와 동일하다. 이와 같이 제 3 계층이 지정됨에 따라, 압축된 신경 네트워크의 제 3 계층의 출력은 제 1 계층의 출력을 모방할 수도 있다. 프로세스는, 제 2 및 제 3 계층을 포함하는 압축된 신경 네트워크의 조합된 메모리 풋프린트는 비압축된 신경 네트워크의 메모리 풋프린트 (M) 보다 더 작도록, (r) 로 나타낼 수도 있는 제 2 계층의 크기를 추가로 지정할 수도 있다. 대안적으로, 또는 메모리 풋프린트에 있어서의 감소에 추가하여, 제 2 계층의 크기는 압축된 신경 네트워크의 추론 시간이 비압축된 신경 네트워크의 추론 시간 (T) 보다 더 작을 수도 있도록 선택될 수도 있다.

[0165] 블록 (1306) 에서, 프로세스는 신경 네트워크의 대체 계층들의 파라미터들을 초기화한다. 하나의 예시적인 구성에서, 압축된 신경 네트워크의 제 2 계층에 투영하는 가중치 매트릭스 (W1), 및 압축된 신경 네트워크의 제 2 계층으로부터 제 3 계층으로 투영하는 가중치 매트릭스 (W2) 는 랜덤 값들로 초기화될 수도 있다. 일부 양태들에서, 가중화 파라미터들의 초기화는 교차 최소화 프로세스를 수반할 수도 있다. 일부 양태들에서, 파라미터들의 초기화는 특이 값 분해를 수반할 수도 있다. 가중치들이 방법의 추가적인 양태들에 따른 추가의 수정들을 받을 수도 있으므로, 가중치들의 초기 세팅이 초기화로서 지칭될 수도 있다.

[0166] 블록 (1308) 에서, 프로세스는 압축된 네트워크의 압축된 계층들 사이에 비선형성을 삽입한다. 예를 들어, 프로세스는 압축된 신경 네트워크의 제 2 계층에서 비선형성을 삽입할 수도 있다. 일부 양태들에서, 프로세스는 비선형 활성화 함수를 압축된 계층들의 뉴런들에 적용함으로써 비선형성을 삽입한다. 비선형 활성화 함수는 교정기, 절대 값 함수, 쌍곡선 탄젠트 함수, 시그모이드 함수, 또는 다른 비선형 활성화 함수를 포함할 수도 있다.

[0167] 블록 (1310) 에서, 프로세스는 계층들 중의 하나 이상에서 가중치 값들을 업데이트함으로써 압축된 네트워크를 미세-튜닝한다. 미세-튜닝은 역 전파 또는 다른 표준 프로세스들을 이용하여 수행될 수도 있다. 미세-튜닝은 압축된 신경 네트워크에서 모든 가중치 값들을 업데이트함으로써, 또는 압축된 계층들의 서브세트, 비압축된 계층들의 서브세트, 또는 양자의 혼합을 업데이트함으로써 수행될 수도 있다. 미세-튜닝은 훈련 예들을 이용하여 수행될 수도 있다. 훈련 예들은 원래의 신경 네트워크를 교육하기 위하여 이용된 것들과 동일할 수도 있거나, 예들의 새로운 세트, 또는 양자의 혼합을 포함할 수도 있다. 예시적인 구성에서, 미세-튜닝은 압축된 신경 네트워크의 제 2 및 제 3 계층들에 대해 수행될 수도 있다. 결국, 신경 네트워크에서의 가중치들의 전부는 태스크 성능의 더 높은 레벨들이 달성될 수도 있도록, 블록 (1312) 에서 조절될 수도 있다.

[0168] 미세-튜닝된, 압축된 신경 네트워크는 예를 들어, 태스크 성능 (P) 을 포함하는 몇몇 차원들을 따라 특성화될 수도 있다. 압축된 신경 네트워크에 대한 태스크 성능이 수용가능한 임계치를 초과할 경우, 프로세스는 종결될 수도 있다. 압축된 신경 네트워크에 대한 태스크 성능이 수용가능한 임계치 미만일 경우 (1314: 아니오), 압축된 신경 네트워크에서의 계층의 크기일 수도 있는 (r) 의 값은 블록 (1316) 에서 증가될 수도 있다. 그렇지 않을 경우 (1314: 예) 에는, 프로세스가 블록 (1318) 에서 종료된다. (r) 의 새로운 값은 메모리 풋프린트 (M) 및/또는 추론 시간 (T) 이 비압축된 신경 네트워크의 그것보다 더 작도록 선택될 수도 있다. 파라미터들을 초기화하고, 비선형성을 삽입하고, 미세-튜닝하고, 성능을 특성화하는 단계들은 수용가능한 태스크 성능에 도달될 때까지 반복될 수도 있다.

[0169] 방법 (1300) 은 수용가능한 태스크 성능이 획득될 때까지, 압축된 계층의 크기 (r) 가 작은 값으로부터 더 큰 값으로 증가될 수도 있는 프로세스를 도시한다. 본 개시물의 양태들에 따르면, (r) 의 값은 압축된 신경 네트워크의 태스크 성능이 수용가능한 태스크 성능을 획득하기 쉽도록, 초기에 보수적으로 선택될 수도 있다. 이에 따라, 방법은 메모리 풋프린트 또는 추론 시간 장점들이 실현될 수도 있도록, 비압축된 신경 네트워크 대신에 이용될 수도 있는 압축된 신경 네트워크를 생성할 수도 있다. 그 다음으로, 방법은 수용가능한 태스크 성능이 획득될 수 없을 때까지, (r) 의 이러한 점점 더 작은 값들을 동작시키는 것을 계속할 수도 있다. 이런 식으로, 방법은 방법의 실행 동안에 압축 레벨 증가를 제공할 수도 있다.

[0170] 위에서 설명된 방법들의 다양한 동작들은 대응하는 기능들을 수행할 수 있는 임의의 적당한 수단에 의해 수행될 수도 있다. 수단은 회로, 애플리케이션 특정 집적 회로 (application specific integrated circuit; ASIC), 또는 프로세서를 포함하지만, 이것으로 제한되지는 않는 다양한 하드웨어 및/또는 소프트웨어 컴포넌트 (들) 및/또는 모듈 (들) 을 포함할 수도 있다. 일반적으로, 도면들에서 예시된 동작들이 있을 경우, 그 동작들은 유사하게 넘버링되는 상응하는 대응 기능식 컴포넌트들을 가질 수도 있다.

[0171] 본원에서 이용된 바와 같이, 용어 "결정" 은 폭넓게 다양한 액션 (action) 들을 망라한다. 예를 들어, "결정" 은 계산, 컴퓨팅, 프로세싱, 유도, 조사, 룩업 (look up) (예컨대, 테이블, 데이터베이스 또는 또 다른 데이터 구조에서의 룩업), 확인 등등을 포함할 수도 있다. 추가적으로, "결정" 은 수신 (예컨대, 정보를 수신하는 것), 액세스 (예컨대, 메모리에서 데이터를 액세스하는 것) 등등을 포함할 수 있다. 또한, "결정" 은 해결, 선택, 선정, 확립 등등을 포함할 수도 있다.

[0172] 본원에서 이용된 바와 같이, 항목들의 리스트 중의 "적어도 하나" 를 지칭하는 어구는 단일 부재들을 포함하는 그러한 항목들의 임의의 조합을 지칭한다. 예로서, "a, b, 또는 c 중의 적어도 하나" 는 a, b, c, a-b, a-c, b-c, 및 a-b-c 를 커버하도록 의도된다.

[0173] 본 개시물과 관련하여 설명된 다양한 예시적인 논리적 블록들, 모듈들, 및 회로들은 범용 프로세서, 디지털 신호 프로세서 (digital signal processor; DSP), 애플리케이션 특정 집적 회로 (ASIC), 필드 프로그래밍가능한 게이트 어레이 신호 (field programmable gate array signal; FPGA) 또는 다른 프로그래밍가능한 로직 디바이스 (PLD), 별개의 게이트 또는 트랜지스터 로직, 별개의 하드웨어 컴포넌트들, 또는 본원에서 설명된 기능들을 수행하도록 설계된 그 임의의 조합으로 구현되거나 수행될 수도 있다. 범용 프로세서는 마이크로프로세서일 수도 있지만, 대안적으로, 프로세서는 임의의 상업적으로 입수가능한 프로세서, 제어기, 마이크로제어기 또는 상태 머신일 수도 있다. 프로세서는 또한, 컴퓨팅 디바이스들의 조합, 예컨대, DSP 및 마이크로프로세서, 복수의 마이크로프로세서들, DSP 코어와 함께 하나 이상의 마이크로프로세서들, 또는 임의의 다른 이러한 구성의 조합으로서 구현될 수도 있다.

[0174] 본 개시물과 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로, 프로세서에 의해 실행된 소프트웨어 모듈로, 또는 이 둘의 조합으로 직접 구체화될 수도 있다. 소프트웨어 모듈은 당해 분야에서 알려져 있는 저장 매체의 임의의 형태로 상주할 수도 있다. 이용될 수도 있는 저장 매체들의 일부 예들은 랜덤 액세스 메모리 (random access memory; RAM), 판독 전용 메모리 (read only memory; ROM), 플래시 메모리, 소거가능 프로그래밍가능 판독-전용 메모리 (erasable programmable read-only memory; EPROM), 전기적 소거가능 프로그래밍가능 판독-전용 메모리 (electrically erasable programmable read-only memory; EEPROM), 레지스터들, 분리가능한 디스크, CD-ROM 등등을 포함한다. 소프트웨어 모듈은 단일 명령 또는 다수의 명령들을 포함할 수도 있고, 몇몇 상이한 코드 세그먼트들 상에서, 상이한 프로그램 사이에서, 그리고 다수의 저장 매체들엘 걸쳐 분산될 수도 있다. 저장 매체는 프로세서가 저장 매체로부터 정보를 판독할 수 있고 정보를 저장 매체에 기록할 수 있도록 프로세서에 결합될 수도 있다. 대안적으로, 저장 매체는 프로세서에 일체적일 수도 있다.

[0175] 본원에서 개시된 방법들은 설명된 방법을 달성하기 위한 하나 이상의 단계들 또는 액션들을 포함한다. 방법 단계들 및/또는 액션들은 청구항들의 범위로부터 이탈하지 않으면서 서로 상호 교환될 수도 있다. 다시 말해서, 단계들 또는 액션들의 특정 순서가 특정되지 않으면, 특정 단계들 및/또는 액션들의 순서 및/또는 이용은 청구항들의 범위로부터 이탈하지 않으면서 수정될 수도 있다.

[0176] 설명된 기능들은 하드웨어, 소프트웨어, 펌웨어, 또는 그 임의의 조합으로 구현될 수도 있다. 하드웨어로 구현될 경우, 일 예의 하드웨어 구성은 디바이스에서의 프로세싱 시스템을 포함할 수도 있다. 프로세싱 시스템은 버스 아키텍처로 구현될 수도 있다. 버스는 프로세싱 시스템의 특정 애플리케이션 및 전체적인 설계 제약들에 따라 임의의 수의 상호접속하는 버스들 및 브릿지 (bridge) 들을 포함할 수도 있다. 버스는 프로세서, 머신 판독가능 매체들, 및 버스 인터페이스를 포함하는 다양한 회로들을 함께 연결할 수도 있다. 버

스 인터페이스는 그 중에서도, 네트워크 어댑터를 버스를 통해 프로세싱 시스템에 접속하기 위하여 이용될 수도 있다. 네트워크 어댑터는 신호 프로세싱 기능들을 구현하기 위하여 이용될 수도 있다. 어떤 양태들에 대하여, 사용자 인터페이스 (예컨대, 키패드, 디스플레이, 마우스, 조이스틱 등) 는 또한 버스에 접속될 수도 있다. 버스는 또한, 당해 분야에서 잘 알려져 있고, 그러므로, 더 이상 설명되지 않을 타이밍 소스들, 주변 기기들, 전압 레귤레이터들, 및 전력 관리 회로들 등등과 같은 다양한 다른 회로들을 연결할 수도 있다.

[0177] 프로세서는 버스를 관리하는 것과, 머신 판독가능 매체들 상에서 저장된 소프트웨어의 실행을 포함하는 일반적인 프로세싱을 담당할 수도 있다. 프로세서는 하나 이상의 범용 및/또는 특수-목적 프로세서들로 구현될 수도 있다. 예들은 마이크로프로세서들, 마이크로제어기들, DSP 프로세서들, 및 소프트웨어를 실행할 수 있는 다른 회로부를 포함한다. 소프트웨어는 소프트웨어, 펌웨어, 미들웨어, 마이크로코드, 하드웨어 설명 언어, 또는 그 외의 것으로서 지칭되든지 간에, 명령들, 데이터, 또는 그 임의의 조합을 의미하도록 폭넓게 해석될 것이다. 머신 판독가능 매체들은 예로서, 랜덤 액세스 메모리 (RAM), 플래시 메모리, 판독 전용 메모리 (ROM), 프로그래밍가능 판독-전용 메모리 (programmable read-only memory; PROM), 소거가능 프로그래밍가능 판독-전용 메모리 (EPROM), 전기적 소거가능 프로그래밍가능 판독-전용 메모리 (EEPROM), 레지스터들, 자기 디스크들, 광학 디스크들, 하드 드라이브들, 또는 임의의 다른 적당한 저장 매체, 또는 그 임의의 조합을 포함할 수도 있다. 머신 판독가능 매체들은 컴퓨터-프로그램 제품에 담길 수도 있다. 컴퓨터-프로그램 제품은 패키징 재료들을 포함할 수도 있다.

[0178] 하드웨어 구현예에서, 머신 판독가능 매체들은 프로세서로부터 분리된 프로세싱 시스템의 일부일 수도 있다. 그러나, 당해 분야의 당업자들이 용이하게 인식하는 바와 같이, 머신 판독가능 매체들 또는 그 임의의 부분은 프로세싱 시스템의 외부에 있을 수도 있다. 예로서, 머신 판독가능 매체들은 송신 라인, 데이터에 의해 변조된 반송파, 및/또는 디바이스로부터 분리된 컴퓨터 제품을 포함할 수도 있고, 그 전부는 버스 인터페이스를 통해 프로세서에 의해 액세스될 수도 있다. 대안적으로 또는 추가적으로, 머신 판독가능 매체들 또는 그 임의의 부분은 캐시 및/또는 일반적인 레지스터 파일들에서 그러한 바와 같이, 프로세서 내로 통합될 수도 있다. 논의된 다양한 컴포넌트들은 로컬 컴포넌트와 같이, 특정 로케이션을 가지는 것으로서 설명될 수도 있지만, 그것들은 또한, 분산된 컴퓨팅 시스템의 일부로서 구성되는 어떤 컴포넌트들과 같이, 다양한 방법들로 구성될 수도 있다.

[0179] 프로세싱 시스템은 외부 버스 아키텍처를 통해 다른 지원 회로부와 함께 모두 연결된, 프로세서 기능성을 제공하는 하나 이상의 마이크로프로세서들 및 머신 판독가능 매체들의 적어도 부분을 제공하는 외부 메모리를 갖는 범용 프로세싱 시스템으로서 구성될 수도 있다. 대안적으로, 프로세싱 시스템은 뉴런 모델들 본원에서 설명된 신경 시스템들의 모델들을 구현하기 위한 하나 이상의 뉴로모픽 프로세서 (neuromorphic processor) 들을 포함할 수도 있다. 또 다른 대안으로서, 프로세싱 시스템은 프로세서, 버스 인터페이스, 사용자 인터페이스, 지원 회로부, 및 단일 칩 내로 통합된 머신 판독가능 매체들의 적어도 부분을 갖는 애플리케이션 특정 집적 회로 (ASIC), 또는 하나 이상의 필드 프로그래밍가능 게이트 어레이 (FPGA) 들, 프로그래밍가능 로직 디바이스 (PLD) 들, 제어기들, 상태 머신들, 게이팅된 로직, 별개의 하드웨어 컴포넌트들, 또는 임의의 다른 적당한 회로부, 또는 이 개시물의 전반에 걸쳐 설명된 다양한 기능성을 수행할 수 있는 회로들의 임의의 조합으로 구현될 수도 있다. 당해 분야의 당업자들은 특정한 애플리케이션 및 전체적인 시스템에 부과된 전체적인 설계 제약들에 따라 프로세싱 시스템을 위한 설명된 기능성을 어떻게 최상으로 구현할 것인지를 인식할 것이다.

[0180] 머신 판독가능 매체들은 다수의 소프트웨어 모듈들을 포함할 수도 있다. 소프트웨어 모듈들은, 프로세서에 의해 실행될 때, 프로세싱 시스템으로 하여금, 다양한 기능들을 수행하게 하는 명령들을 포함한다. 소프트웨어 모듈들은 송신 모듈 및 수신 모듈을 포함할 수도 있다. 각각의 소프트웨어 모듈은 단일 저장 디바이스에서 상주할 수도 있거나, 다수의 저장 디바이스들에 걸쳐 분산될 수도 있다. 예로서, 소프트웨어 모듈은 트리거링 이벤트가 발생할 때에 하드 드라이브로부터 RAM 으로 로딩될 수도 있다. 소프트웨어 모듈의 실행 동안, 프로세서는 액세스 속력을 증가시키기 위하여 명령들의 일부를 캐시로 로딩할 수도 있다. 그 다음으로, 하나 이상의 캐시 라인들은 프로세서에 의한 실행을 위하여 일반적인 레지스터 파일로 로딩될 수도 있다. 이하의 소프트웨어 모듈의 기능성을 지칭할 때, 이러한 기능성은 그 소프트웨어 모듈로부터의 명령들을 실행할 때에 프로세서에 의해 구현된다는 것이 이해될 것이다. 또한, 본 개시물의 양태들은 프로세서, 컴퓨터, 머신, 또는 이러한 양태들을 구현하는 다른 시스템에 대한 개선들로 귀착된다는 것이 인식되어야 한다.

[0181] 소프트웨어로 구현될 경우, 기능들은 하나 이상의 명령들 또는 코드로서, 컴퓨터 판독가능 매체 상에 저장되거나, 컴퓨터 판독가능 매체 상에서 송신될 수도 있다. 컴퓨터 판독가능 매체들은, 하나의 장소로부터 또 다른 장소까지의 컴퓨터 프로그램의 전송을 가능하게 하는 임의의 매체를 포함하는 컴퓨터 저장 매체들 및 통신

매체들의 양자를 포함한다. 저장 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 이용가능한 매체일 수도 있다. 제한이 아닌 예로서, 이러한 컴퓨터 판독가능한 매체들은 RAM, ROM, EEPROM, CD-ROM 또는 다른 광학 디스크 저장, 자기 디스크 저장, 또는 다른 자기 저장 디바이스들, 또는 명령들 또는 데이터 구조들의 형태로 바람직한 프로그램 코드를 반송 또는 저장하기 위해 이용될 수 있으며 컴퓨터에 의해 액세스될 수 있는 임의의 다른 매체를 포함할 수 있다. 게다가, 임의의 접속은 컴퓨터 판독가능 매체로 적절하게 칭해진다. 예를 들어, 동축 케이블, 광섬유 케이블, 트위스트 페어(twisted pair), 디지털 가입자 회선(digital subscriber line; DSL), 또는 적외선, 라디오(radio), 및 마이크로파(microwave) 와 같은 무선 기술들을 이용하여, 소프트웨어가 웹사이트, 서버, 또는 다른 원격 소스로부터 송신될 경우, 동축 케이블, 광섬유 케이블, 트위스트 페어, DSL, 또는 적외선, 라디오, 및 마이크로파와 같은 무선 기술들은 매체의 정의 내에 포함된다. 본원에서 이용된 바와 같은 디스크 (disk) 및 디스크 (disc) 는 컴팩트 디스크 (compact disc; CD), 레이저 디스크 (laser disc), 광학 디스크 (optical disc), 디지털 다기능 디스크 (digital versatile disc; DVD), 플로피 디스크 (floppy disk) 및 BLu-ray® 디스크 (disc) 를 포함하고, 여기서, 디스크 (disk) 들은 통상 데이터를 자기적으로 재생하는 반면, 디스크 (disc) 들은 데이터를 레이저들로 광학적으로 재생한다. 이에 따라, 일부 양태들에서, 컴퓨터 판독가능 매체들은 비-일시적인 컴퓨터 판독가능 매체들 (예컨대, 유형의 (tangible) 매체들) 을 포함할 수도 있다. 게다가, 다른 양태들에 대하여, 컴퓨터 판독가능 매체들은 일시적 컴퓨터 판독가능 매체들 (예컨대, 신호) 을 포함할 수도 있다. 상기의 조합들은 또한, 컴퓨터 판독가능 매체들의 범위 내에 포함되어야 한다.

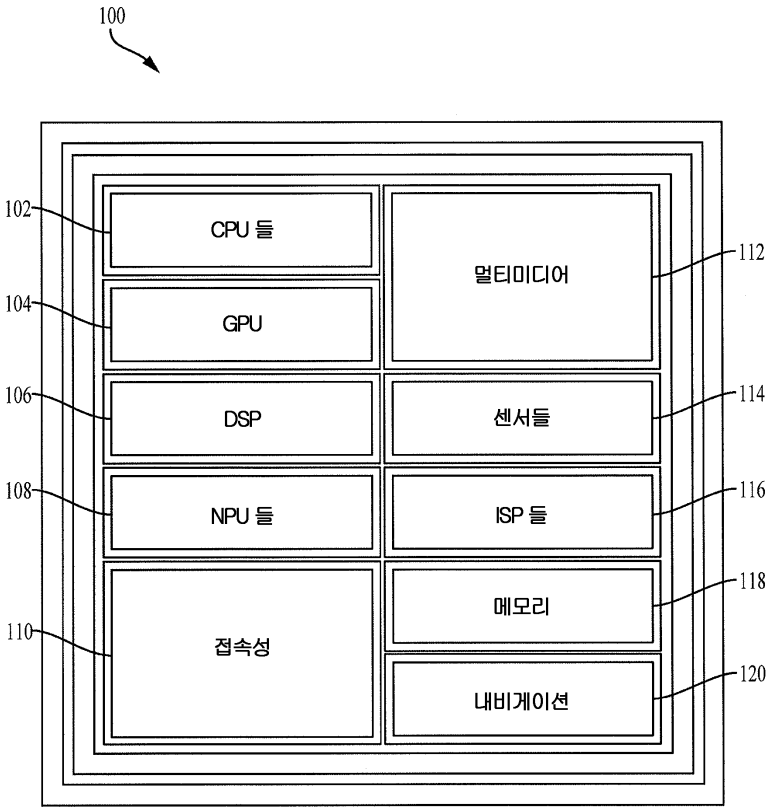
[0182] 이에 따라, 어떤 양태들은 본원에서 제시된 동작들을 수행하기 위한 컴퓨터 프로그램 제품을 포함할 수도 있다. 예를 들어, 이러한 컴퓨터 프로그램 제품은 명령들을 그 위에 저장한 (및/또는 인코딩한) 컴퓨터 판독가능 매체를 포함할 수도 있고, 명령들은 본원에서 설명된 동작들을 수행하기 위하여 하나 이상의 프로세서들에 의해 실행가능할 수도 있다. 어떤 양태들에 대하여, 컴퓨터 프로그램 제품은 패키징 재료를 포함할 수도 있다.

[0183] 또한, 본원에서 설명된 방법들 및 기법들을 수행하기 위한 모듈들 및/또는 다른 적절한 수단은 적용가능한 바와 같은 사용자 단말 및/또는 기지국에 의해 다운로딩될 수 있고 및/또는 이와 다르게 획득될 수 있다는 것이 인식되어야 한다. 예를 들어, 이러한 디바이스는 본원에서 설명된 방법들을 수행하기 위한 수단의 전달을 용이하게 하기 위하여 서버에 결합될 수 있다. 대안적으로, 본원에서 설명된 다양한 방법들은 저장 수단 (예를 들어, RAM, ROM, 물리적 저장 매체 예컨대, 컴팩트 디스크 (CD) 또는 플로피 디스크, 등) 을 통해 제공될 수 있어서, 사용자 단말 및/또는 기지국은 저장 수단을 디바이스에 결합 또는 제공 시에 다양한 방법들을 획득할 수 있다. 게다가, 본원에서 설명된 방법들 및 기법들을 디바이스에 제공하기 위한 임의의 다른 적당한 기법이 사용될 수 있다.

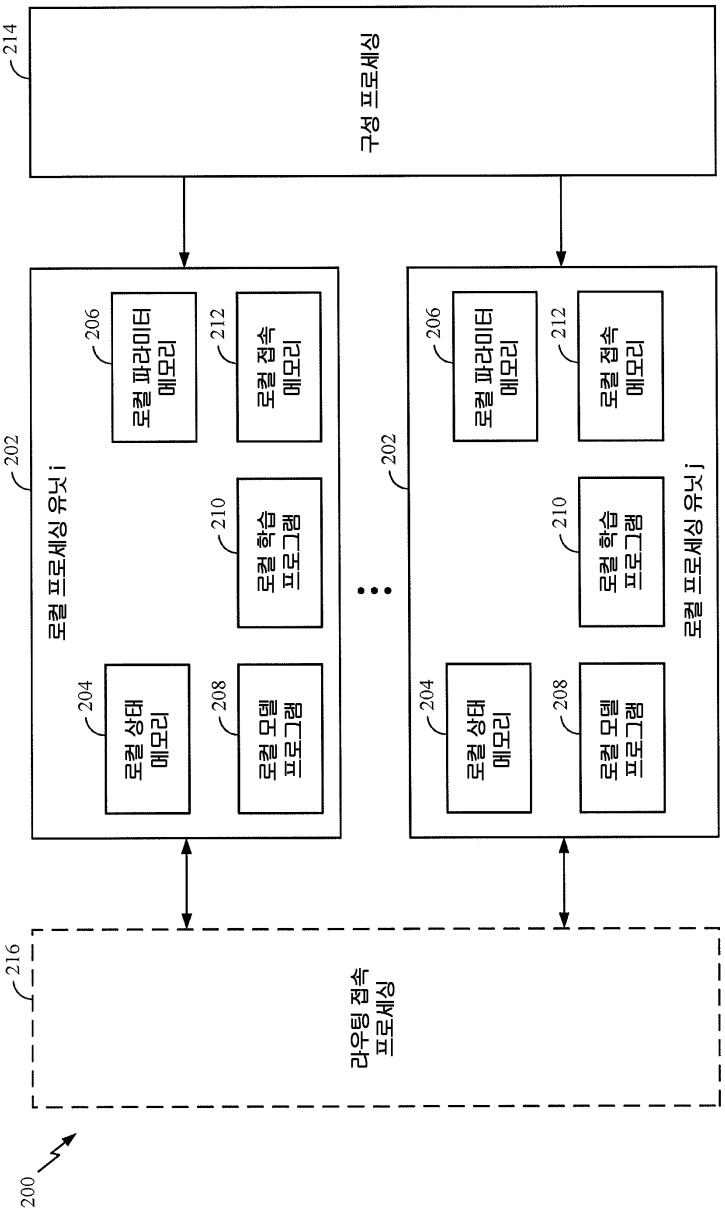
[0184] 청구항들은 위에서 예시된 정확한 구성 및 컴폰너트들에 제한되지 않는다는 것이 이해되어야 한다. 다양한 수정들, 변경들 및 변동들은 청구항들의 범위로부터 이탈하지 않으면서, 위에서 설명된 방법들 및 장치의 배열, 동작 및 세부사항들에서 행해질 수도 있다.

도면

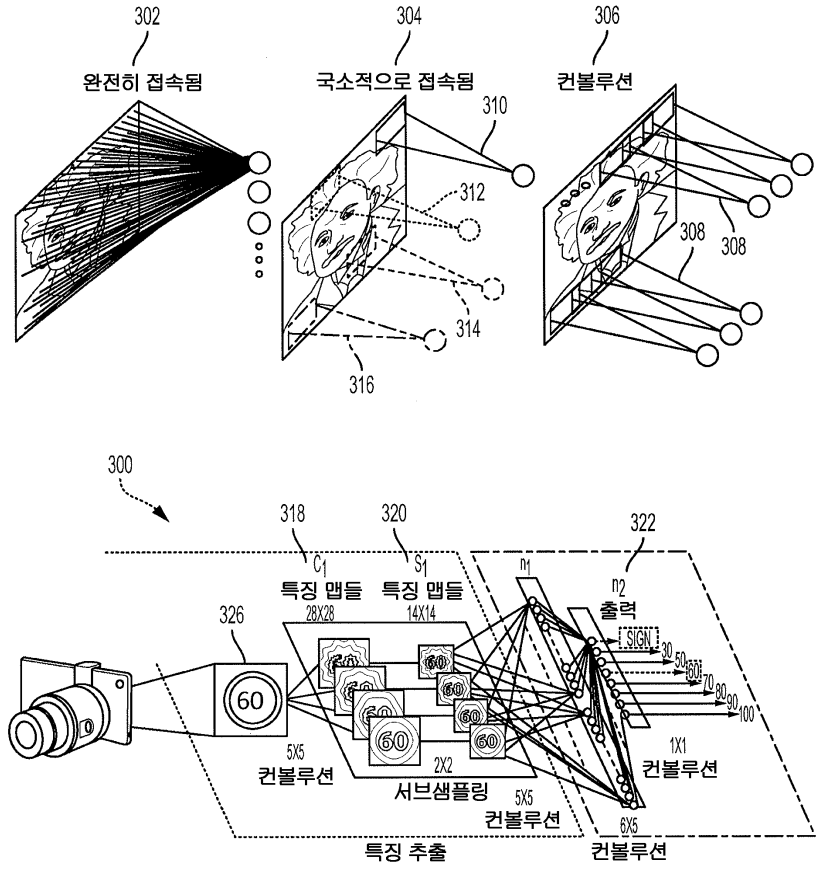
도면1


100
102
CPU 들
104
GPU
106
DSP
108
NPU 들
110
접속성
112
멀티미디어
114
센서들
116
ISP 들
118
메모리
120
내비게이션

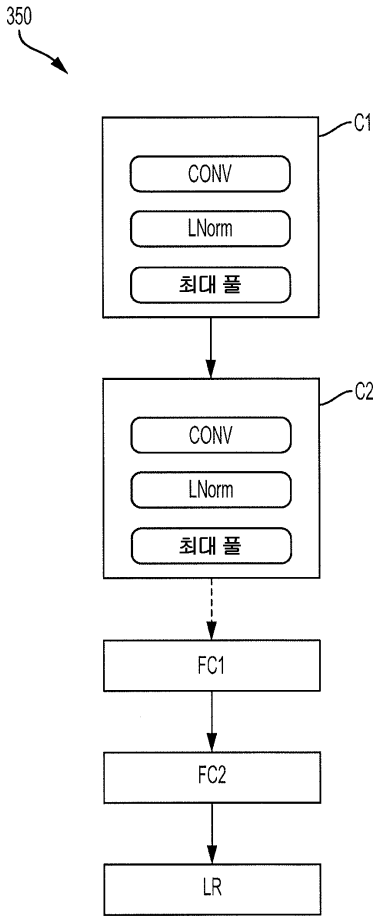
도면2


214
구성 프로세싱
202
로컬 프로세싱 유닛 i
206
로컬 파라미터 메모리
212
로컬 접속 메모리
210
로컬 학습 프로그램
204
로컬 상태 메모리
208
로컬 모델 프로그램
202
로컬 프로세싱 유닛 j
206
로컬 파라미터 메모리
212
로컬 접속 메모리
210
로컬 학습 프로그램
204
로컬 상태 메모리
208
로컬 모델 프로그램
216
라우팅 접속 프로세싱
200

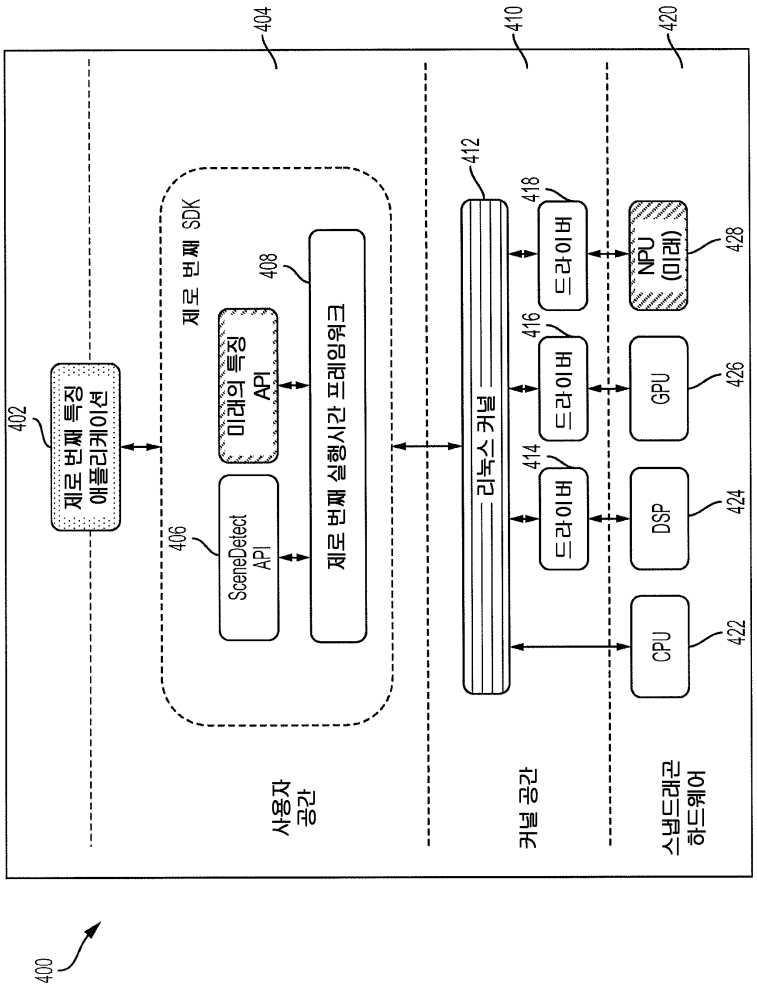
도면3a


302
304
306
완전히 접속됨
국소적으로 접속됨
컨볼루션
310
312
314
316
308
308
300
318
320
322
C1
S1
n1
n2
특징 맵들
특징 맵들
출력
28X28
14X14
326
SIGN
30
50
60
70
80
90
100
60
5X5
컨볼루션
2X2
서브샘플링
5X5
컨볼루션
1X1
컨볼루션
6X5
컨볼루션
특징 추출

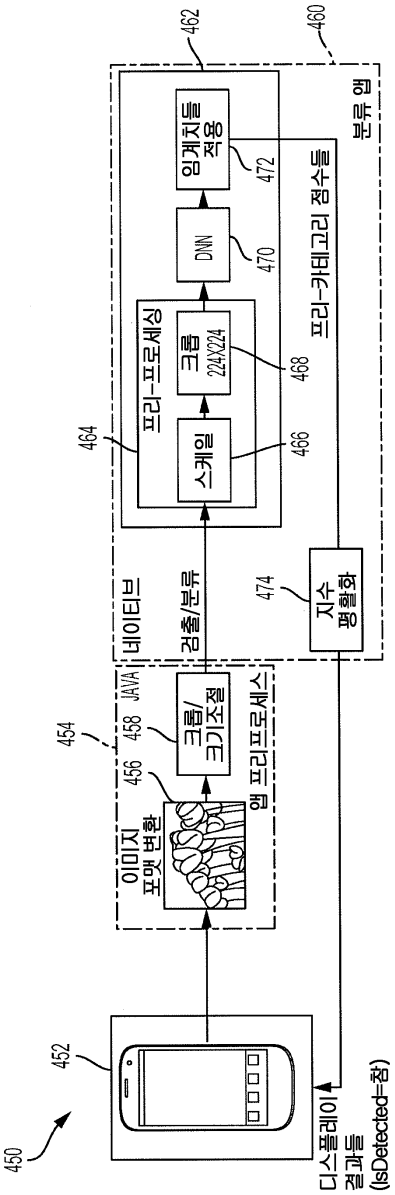
도면3b



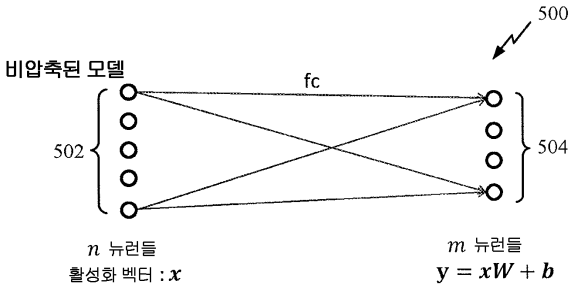
도면4a


400
404
410
420
402
제로 번째 특징 애플리케이션
제로 번째 SDK
406
SceneDetect API
408
미래의 특징 API
제로 번째 실행시간 프레임워크
412
리눅스 커널
414
드라이버
416
드라이버
418
드라이버
CPU
422
DSP
424
GPU
426
NPU (미래)
428
사용자 공간
커널 공간
스냅드래곤 하드웨어

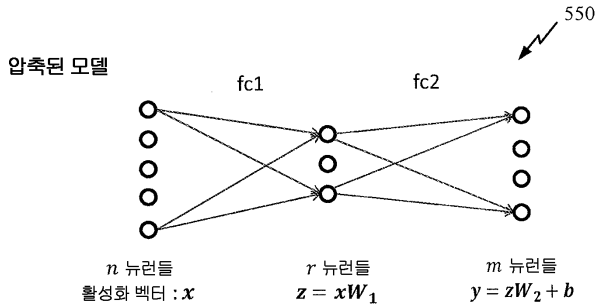
도면4b


462
460
분류 앱
임계치들 적용
472
DNN
470
프리-카테고리 점수들
프리-프로세싱
크롭 224X224
468
스케일
466
464
네이티브
검출/분류
474
지수 평활화
JAVA
454
458
크롭/ 크기조절
456
앱 프리프로세스
이미지 포맷 변환
452
450
디스플레이 결과들 (isDetected=참)

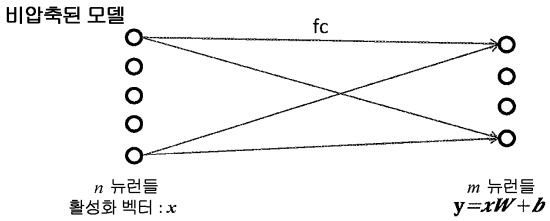
도면5a


500
비압축된 모델
fc
502
504
n 뉴런들
활성화 벡터 : x
m 뉴런들
y = xW + b

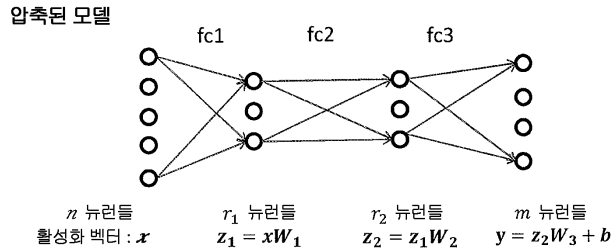
도면5b



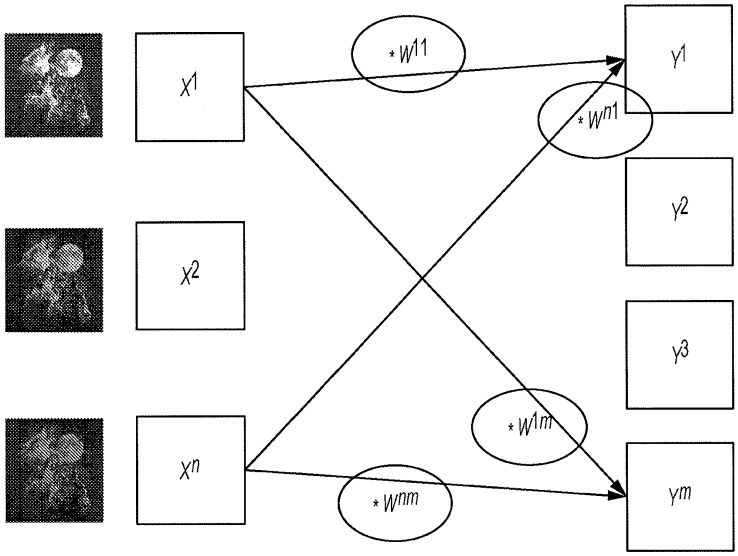
도면6a



도면6b

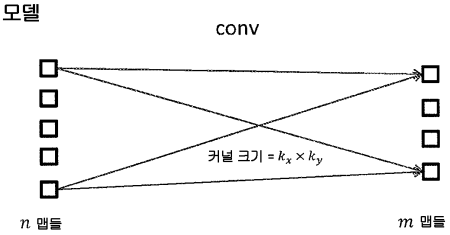


도면7



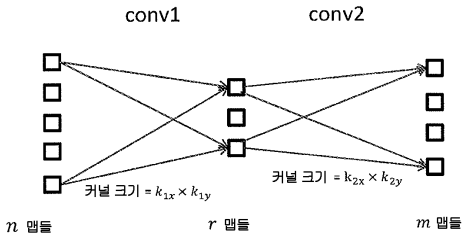
도면8a

비압축된 모델



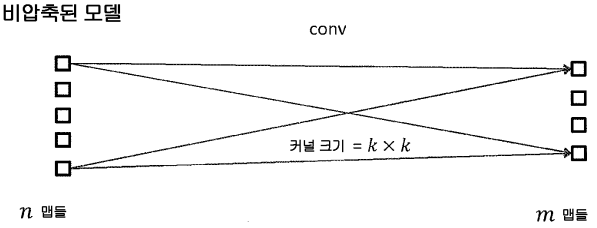
도면8b

압축된 모델



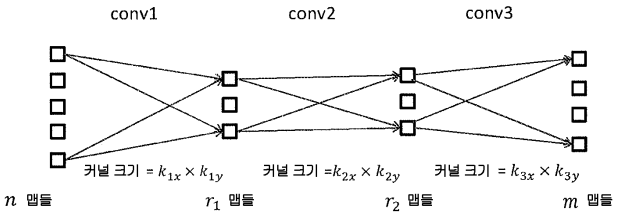
도면9a

비압축된 모델

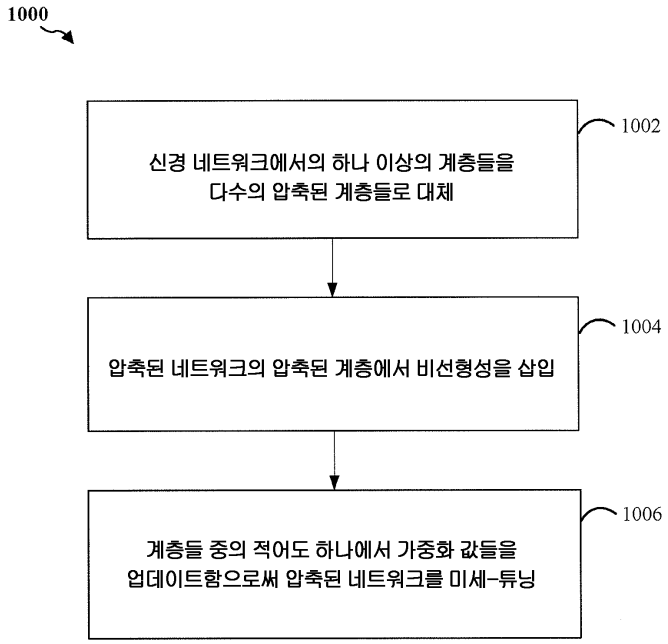


도면9b

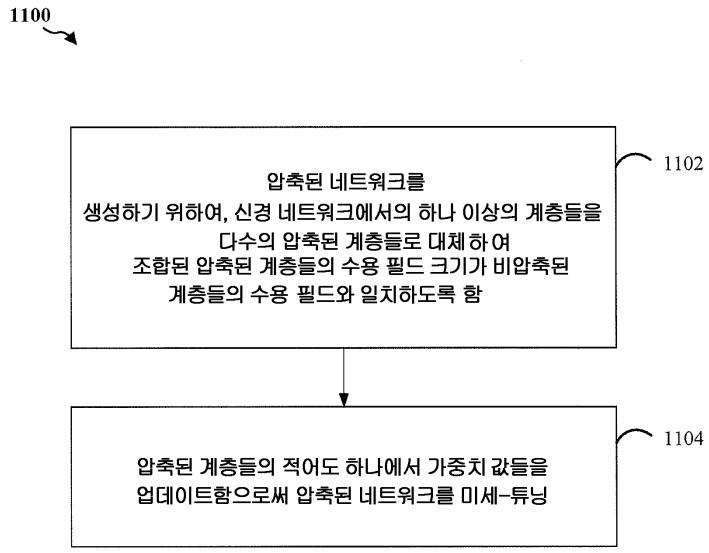
압축된 모델



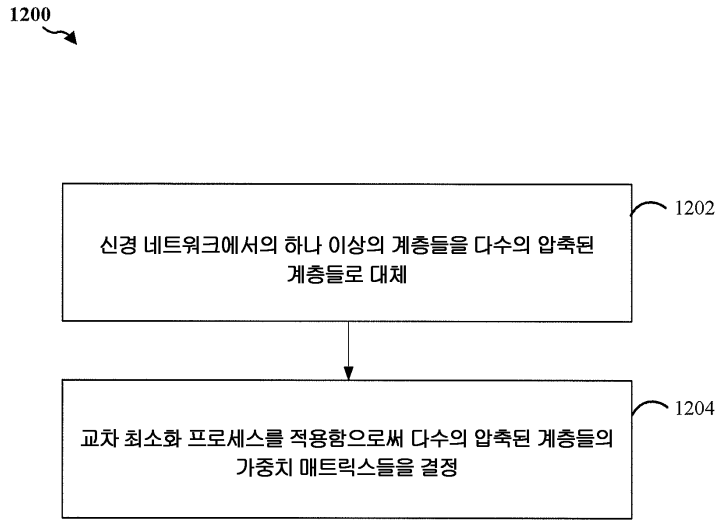
도면10


1000
신경 네트워크에서의 하나 이상의 계층들을
다수의 압축된 계층들로 대체
1002
압축된 네트워크의 압축된 계층에서 비선형성을 삽입
1004
계층들 중의 적어도 하나에서 가중화 값들을
업데이트함으로써 압축된 네트워크를 미세-튜닝
1006

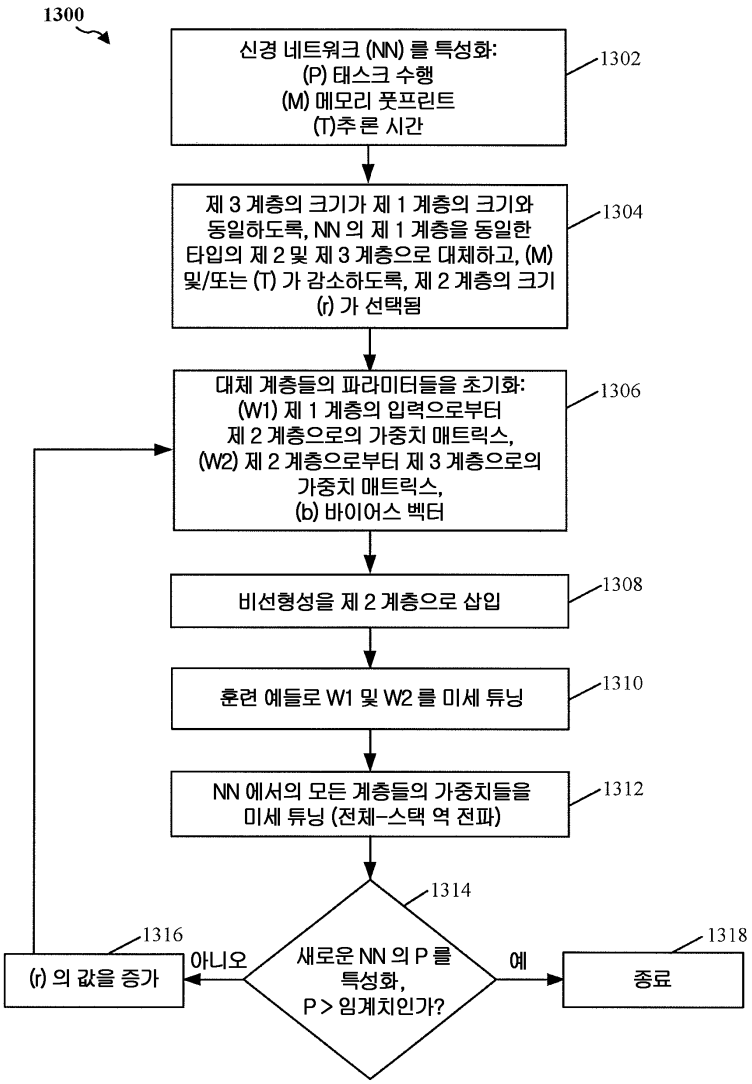
도면11


1100
압축된 네트워크를
생성하기 위하여, 신경 네트워크에서의 하나 이상의 계층들을
다수의 압축된 계층들로 대체하여
조합된 압축된 계층들의 수용 필드 크기가 비압축된
계층들의 수용 필드와 일치하도록 함
1102
압축된 계층들의 적어도 하나에서 가중치 값들을
업데이트함으로써 압축된 네트워크를 미세-튜닝
1104

도면12


1200
신경 네트워크에서의 하나 이상의 계층들을 다수의 압축된 계층들로 대체
1202
교차 최소화 프로세스를 적용함으로써 다수의 압축된 계층들의 가중치 매트릭스들을 결정
1204

도면13


1300
신경 네트워크 (NN) 를 특성화:
(P) 태스크 수행
(M) 메모리 풋프린트
(T)추론 시간
1302
제 3 계층의 크기가 제 1 계층의 크기와 동일하도록, NN 의 제 1 계층을 동일한 타입의 제 2 및 제 3 계층으로 대체하고, (M) 및/또는 (T) 가 감소하도록, 제 2 계층의 크기 (r) 가 선택됨
1304
대체 계층들의 파라미터들을 초기화:
(W1) 제 1 계층의 입력으로부터 제 2 계층으로의 가중치 매트릭스,
(W2) 제 2 계층으로부터 제 3 계층으로의 가중치 매트릭스,
(b) 바이어스 벡터
1306
비선형성을 제 2 계층으로 삽입
1308
훈련 예들로 W1 및 W2 를 미세 튜닝
1310
NN 에서의 모든 계층들의 가중치들을 미세 튜닝 (전체-스택 역 전파)
1312
새로운 NN 의 P 를 특성화,
P > 임계치인가?
1314
아니오
예
(r) 의 값을 증가
1316
종료
1318