

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2021-0069506
(43) 공개일자 2021년06월11일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 20/10 (2019.01)
G06N 3/04 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 20/10 (2019.01)
(21) 출원번호 10-2019-0159412
(22) 출원일자 2019년12월03일
심사청구일자 2019년12월03일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
배성호
경기도 용인시 기흥구 예현로 15 서그내마을에스케이아파트 102동 501호
이강호
경기도 수원시 영통구 영일로 16-8, 203호 (영통동)
정준현
서울특별시 서초구 남부순환로350길 29-10, B01호 (양재동)
(74) 대리인
특허법인지원

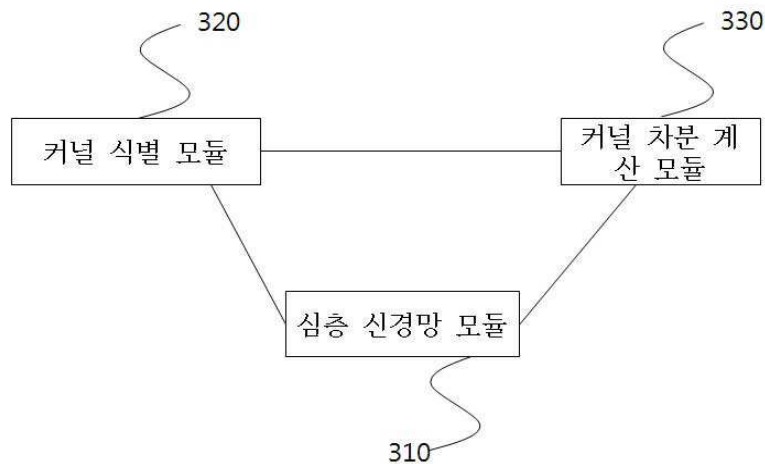
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 심층 신경망 시스템을 포함하는 전자 장치 및 그 동작 방법

(57) 요약

본 출원의 일 실시예에 따르면, 획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 데이터 베이스; 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 설정된(configured to, 또는 implemented to) 커널 식별 모듈; 및 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하도록 설정된 커널 차분 계산 모듈;을 포함하는, 전자 장치가 제공된다.

대표도 - 도2



(52) CPC특허분류
G06N 3/04 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	10085646
부처명	과학기술정보통신부
과제관리(전문)기관명	한국산업기술평가관리원
연구사업명	전자정보디바이스산업원천기술개발사업(반도체)
연구과제명	3D 맵리스터 어레이 결점 극복을 위한 뉴로모픽 시스템
기 여 율	1/2
과제수행기관명	경희대학교산학협력단
연구기간	2019.06.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	2015-0-00742
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	Grand ICT연구센터 지원사업
연구과제명	라이프 컴패니온쉽 경험을 위한 지능형 인터렉션 융합 연구
기 여 율	1/2
과제수행기관명	성균관대학교 산학협력단
연구기간	2019.01.01 ~ 2019.12.31

명세서

청구범위

청구항 1

획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 데이터 베이스;

상기 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 설정된(configured to, 또는 implemented to) 커널 식별 모듈; 및

상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하도록 설정된 커널 차분 계산 모듈;을 포함하는, 전자 장치.

청구항 2

제 1 항에 있어서,

상기 복수 개의 레이어들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고,

상기 커널 식별 모듈은:

상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하고,

상기 복수 개의 레이어들 중 상기 제 1 레이어의 이전의 추론 순서인 제 2 레이어에 포함된 커널들과 상기 제 1 커널 간의 디스턴스를 계산하고,

상기 제 2 레이어에 포함된 커널들 중 계산된 디스턴스가 가장 작은 커널을 상기 제 2 커널로 식별하도록 더 설정된, 전자 장치.

청구항 3

제 1 항에 있어서,

상기 복수 개의 커널들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고,

상기 커널 식별 모듈은:

상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하고,

상기 제 1 레이어로부터 상기 제 1 커널의 위치를 식별하고,

상기 제 1 레이어의 바로 이전 추론 순서인 상기 제 2 레이어로부터 상기 식별된 제 1 커널의 위치와 동일한 위치의 커널을 상기 제 2 커널로 식별하도록 더 설정된, 전자 장치.

청구항 4

제 3 항에 있어서,

상기 제 1 레이어의 제 1 커널과 상기 제 1 레이어와 바로 이전 추론 순서인 상기 제 2 레이어의 제 1 커널의 위치와 동일한 위치인 제 2 커널과의 가중치의 유사도가, 상기 제 1 커널과 다른 커널과의 가중치의 유사도 보다 큰, 전자 장치.

청구항 5

제 1 항에 있어서,

상기 커널 식별 모듈은, 상기 복수 개의 커널들 중 상기 서로 인접한 상기 제 1 커널 및 상기 제 2 커널 이외의 나머지 커널들 중에서 서로 인접한 커널들을 식별하도록 더 설정되고

상기 커널 차분 계산 모듈은, 상기 서로 인접한 커널들의 제 1 차분들을 계산하고 저장하도록 더 설정된, 전자 장치.

청구항 6

제 5 항에 있어서,

상기 계산된 차분은, 상기 제 1 커널 및 상기 제 2 커널에 관련된 인덱스와 함께 저장되고,

상기 계산된 제 1 차분들은, 상기 제 1 차분들의 각각과 관련된 서로 인접한 커널들의 각각과 관련된 인덱스와 함께 저장되는,

전자 장치.

청구항 7

제 5 항에 있어서

상기 복수 개의 커널들의 가중치의 분산 값 보다 상기 저장된 계산된 차분과 상기 저장된 계산된 제 1 차분들의 가중치의 분산 값이 더 작은 것을 특징으로 하는,

전자 장치.

청구항 8

제 5 항에 있어서,

상기 계산된 차분 및 상기 계산된 제 1 차분들은, 양자화 및 지정된 코딩 방식을 기반으로, 저장되는,

전자 장치.

청구항 9

제 1 항에 있어서,

상기 제 1 레이어의 가중치와 상기 제 2 레이어의 가중치의 차이를 감소시키기 위해 상기 복수 개의 레이어들 간의 손실을 계산하고,

상기 계산된 손실을 목적 함수로 하여, 상기 계산된 손실이 최소화되도록 상기 복수 개의 레이어들에 포함된 커널들의 가중치 값들을 재구성하고,

상기 재구성 이후, 상기 커널들의 가중치 값들의 차분을 저장하도록 설정된, 레이어 손실 계산 모듈;을 더 포함하는, 전자 장치.

청구항 10

획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 단계;
 상기 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하는 단계; 및
 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하는 단계;를 포함하는, 전자 장치의 제어 방법.

청구항 11

제 10 항에 있어서,
 상기 복수 개의 레이어들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고,
 상기 제어 방법은:
 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하는 단계;
 상기 복수 개의 레이어들 중 상기 제 1 레이어의 이전의 추론 순서인 제 2 레이어에 포함된 커널들과 상기 제 1 커널 간의 디스턴스를 계산하는 단계; 및
 상기 제 2 레이어에 포함된 커널들 중 계산된 디스턴스가 가장 작은 커널을 상기 제 2 커널로 식별하는 단계;를 포함하는, 전자 장치의 제어 방법.

청구항 12

제 10 항에 있어서,
 상기 복수 개의 커널들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고,
 상기 제어 방법은:
 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하는 단계;
 상기 제 1 레이어로부터 상기 제 1 커널의 위치를 식별하는 단계; 및
 상기 제 1 레이어의 바로 이전 추론 순서인 상기 제 2 레이어로부터 상기 식별된 제 1 커널의 위치와 동일한 위치의 커널을 상기 제 2 커널로 식별하는 단계;를 포함하는, 전자 장치의 제어 방법.

청구항 13

제 12 항에 있어서,
 상기 제 1 레이어의 제 1 커널과 상기 제 1 레이어와 바로 이전 추론 순서인 상기 제 2 레이어의 제 1 커널의 위치와 동일한 위치인 제 2 커널과의 가중치의 유사도가, 상기 제 1 커널과 다른 커널과의 가중치의 유사도 보다 큰, 전자 장치의 제어 방법.

청구항 14

제 13 항에 있어서,
 상기 커널 식별 모듈은, 상기 복수 개의 커널들 중 상기 서로 인접한 상기 제 1 커널 및 상기 제 2 커널 이외의 나머지 커널들 중에서 서로 인접한 커널들을 식별하는 단계; 및

상기 서로 인접한 커널들의 제 1 차분들을 계산하고 저장하는 단계;를 더 포함하는 전자 장치의 제어 방법.

청구항 15

제 14 항에 있어서,

상기 계산된 차분은, 상기 제 1 커널 및 상기 제 2 커널에 관련된 인덱스와 함께 저장되고,

상기 계산된 제 1 차분들은, 상기 제 1 차분들의 각각과 관련된 서로 인접한 커널들의 각각과 관련된 인덱스와 함께 저장되는, 전자 장치의 제어 방법.

청구항 16

제 14 항에 있어서

상기 복수 개의 커널들의 가중치의 분산 값 보다 상기 저장된 계산된 차분과 상기 저장된 계산된 제 1 차분들의 가중치의 분산 값이 더 작은 것을 특징으로 하는, 전자 장치의 제어 방법.

청구항 17

제 14 항에 있어서,

상기 계산된 차분 및 상기 계산된 제 1 차분들은, 양자화 및 지정된 코딩 방식을 기반으로, 저장되는, 전자 장치의 제어 방법.

청구항 18

제 10 항에 있어서,

상기 제 1 레이어의 가중치와 상기 제 2 레이어의 가중치의 차이를 감소시키기 위해 상기 복수 개의 레이어들 간의 손실을 계산하는 동작;

상기 계산된 손실을 목적 함수로 하여, 상기 계산된 손실이 최소화되도록 상기 복수 개의 레이어들에 포함된 커널들의 가중치 값들을 재구성하는 동작; 및

상기 재구성 이후, 상기 커널들의 가중치 값들의 차분을 저장하도록 하는 동작;을 더 포함하는, 전자 장치의 제어 방법.

청구항 19

제 18 항에 있어서,

상기 재구성 이후 저장된 가중치 값들의 차분의 데이터 양은, 상기 재구성 이전 저장된 가중치 값들의 차분의 데이터 양 보다 적은, 전자 장치의 제어 방법.

청구항 20

제어부; 및

획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델 및 인스트럭션들을 저장하는 메모리;를 포함하고,

상기 인스트럭션들이 실행되는 경우, 상기 인스트럭션들은 상기 제어부가:

상기 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 하고,

상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하도록 하는, 전자 장치.

발명의 설명

기술 분야

- [0001] 본 출원은, 심층 신경망 시스템을 포함하는 전자 장치 및 그 동작 방법에 관한 것으로, 보다 상세하게는 압축률 향상을 위해 인공 신경망 모델의 서로 다른 레이어의 각각으로부터 식별된 커널의 가중치 간 차분에 기반하여 학습 및 추론하는 전자 장치 및 그 동작 방법에 관한 것이다.

배경 기술

- [0003] 영상 처리 등의 다양한 분야에서 심층 신경망 모델에 대한 수요가 증가됨에 따라, 심층 신경망 모델이 제한된 환경(예: 개인 단말 등의 환경에서 심층 신경망 모델이 이용됨)에서 이용되는 경우가 빈번해져, 이 경우 심층 신경망 모델의 구동이 어려워지는 문제점이 발생되고 있다.
- [0004] 따라서, 제한된 환경에서 이용될 수 있도록 높은 압축률을 제공하는 심층 신경망 모델에 대한 수요가 증대 되고 있는 실정이다.

발명의 내용

해결하려는 과제

- [0006] 본 출원의 일 실시예에 따르면, 일 과제는 인공 신경망 모델에 포함된 서로 다른 추론 순서를 가지는 레이어로부터 커널들을 식별하고, 커널들 간 차분을 계산하여 저장함으로써 압축률이 향상된 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법을 제공하는 것에 있다.
- [0007] 본 출원의 일 실시예에 따르면, 다른 과제는 커널들 간 차분의 저장 시 양자화 및 지정된 코딩 방식(예: 엔트로피 코딩 방식)을 기반으로 커널들 간 차분을 저장하여 더 향상된 압축률을 가지는 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법을 제공하는 것에 있다.
- [0008] 본 출원의 일 실시예에 따르면, 또 다른 과제는 커널들의 가중치의 차이를 감소시키도록 하는 목적 함수를 기반으로 학습하여, 더 향상된 압축률을 가지는 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법을 제공하는 것에 있다.
- [0009] 본 출원이 해결하고자 하는 과제가 상술한 과제로 제한되는 것은 아니며, 언급되지 아니한 과제들은 본 명세서 및 첨부된 도면으로부터 본 출원이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

- [0011] 본 출원의 일 실시예에 따르면, 획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 데이터 베이스; 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 설정된(configured to, 또는 implemented to) 커널 식별 모듈; 및 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분 및 참조된 커널(제 2 커널)의 인덱스를 저장하도록 설정된 커널 차분 계산 모듈;을 포함하는, 전자 장치가 제공될 수 있다.
- [0012] 본 출원의 다른 실시예에 따르면, 획득되는 입력 데이터에 대한 추론을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 단계; 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제

1 커널 및 제 2 커널을 식별하는 단계; 및 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하는 단계;를 포함하는, 전자 장치의 제어 방법이 제공될 수 있다.

- [0013] 본 출원의 또 다른 실시예에 따르면, 제어부; 및 획득되는 입력 데이터에 대한 추론을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델 및 인스트럭션들을 저장하는 메모리;를 포함하고, 상기 인스트럭션들이 실행되는 경우, 상기 인스트럭션들은 상기 제어부가: 상기 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 하고, 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분 및 제2 커널(참조되어지는 커널)의 인덱스를 저장하도록 하는, 전자 장치가 제공될 수 있다.
- [0014] 과제의 해결 수단이 상술한 해결 수단들로 제한되는 것은 아니며, 언급되지 아니한 해결 수단들은 본 명세서 및 첨부된 도면으로부터 본 출원이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

발명의 효과

- [0016] 본 출원의 일 실시예에 따르면, 인공 신경망 모델이 포함된 서로 다른 추론 순서를 가지는 레이어로부터 커널들을 식별하고, 커널들 간 차분을 계산하여 저장함으로써 압축률이 향상된 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법이 제공될 수 있다.
- [0017] 본 출원의 일 실시예에 따르면, 커널들 간 차분의 저장 시 양자화 및 지정된 코딩 방식(예: 엔트로피 코딩 방식)을 기반으로 커널들 간 차분을 저장하여 더 향상된 압축률을 가지는 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법이 제공될 수 있다.
- [0018] 본 출원의 일 실시예에 따르면, 커널들의 가중치의 차이를 감소시키도록 하는 목표 함수를 기반으로 학습하여, 더 향상된 압축률을 가지는 인공 신경망 모델을 포함하는 전자 장치 및 동작 방법들이 제공될 수 있다.
- [0019] 다양한 실시예들에 따르면, 효과가 상술한 효과로 제한되는 것은 아니며, 언급되지 아니한 효과들은 본 명세서 및 첨부된 도면으로부터 본 출원이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확히 이해될 수 있을 것이다.

도면의 간단한 설명

- [0021] 도 1은 본 출원의 일 실시예에 따른 인공 신경망 시스템이 구현될 수 있는 전자 장치를 설명하기 위한 도면이다.
- 도 2는 본 출원의 일 실시예에 따른 인공 신경망 시스템에 포함된 모듈들을 설명하기 위한 도면이다.
- 도 3은 본 출원의 일 실시예에 따른 서버 디바이스 및/또는 로컬 디바이스의 구성을 나타내는 블록도이다.
- 도 4는, 본 출원의 일 실시예에 따른 인공 신경망 시스템의 커널간 차분 저장 동작의 일 예를 설명하기 위한 흐름도이다.
- 도 5는 본 출원의 일 실시예에 따른 인공 신경망 시스템의 커널간 차분 저장 동작을 설명하기 위한 도면이다.
- 도 6은 본 출원의 일 실시예에 따른 인공 신경망 시스템의 가중치 차분 계산을 위한 레이어를 식별하는 동작을 구체적으로 설명하기 위한 도면이다.
- 도 7은 본 출원의 일 실시예에 따른 인접 레이어의 가중치들의 관계를 설명하기 위한 도면이다.
- 도 8은 가중치의 분포 및 가중치의 차분의 분포를 나타내는 그래프이다.
- 도 9는 본 출원의 일 실시예에 따른 적어도 하나의 목표 함수(예: inter-layer loss)를 기반으로 학습을 수행하는 인공 신경망 시스템의 동작의 일 예를 설명하기 위한 흐름도이다.
- 도 10은 λ (목적함수에서의 inter-layer loss에 대한 가중치)에 따른 인공 신경망 모델 파라미터의 데이터의 양 및 학습 정확도를 나타내는 그래프이다.

도 11은 본 출원의 일 실시예에 따른 인접 레이어의 가중치들의 관계를 설명하기 위한 도면이다.

도 12는 심층신경망 학습시 inter layer loss가 목적함수에 포함되지 않은 모델과 inter layer loss가 목적함수에 포함된 모델의 가중치에 대한 차분의 분포를 나타내는 그래프이다.

도 13은 본 출원의 일 실시예에 따른 다양한 인공 신경망 모델들의 파라미터의 데이터 양을 나타내는 도면이다.

도 14는 본 출원의 일 실시예에 따른 다양한 인공 신경망 모델들의 파라미터의 데이터 양 대비 테스트 정확도를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0022] 본 출원의 일 실시예에 따르면, 획득되는 입력 데이터에 대한 추론을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 데이터 베이스; 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 설정된(configured to, 또는 implemented to) 커널 식별 모듈; 및 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분 및 제 2 커널의 인덱스를 저장하도록 설정된 커널 차분 계산 모듈;을 포함하는, 전자 장치가 제공될 수 있다.
- [0023] 본 출원의 일 실시예에 따르면, 상기 복수 개의 레이어들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고, 상기 커널 식별 모듈은: 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하고, 상기 복수 개의 레이어들 중 상기 제 1 레이어의 이전의 추론 순서인 제 2 레이어에 포함된 커널들과 상기 제 1 커널 간의 디스턴스를 계산하고, 상기 제 2 레이어에 포함된 커널들 중 계산된 디스턴스가 가장 작은 커널을 상기 제 2 커널로 식별하도록 더 설정된, 전자 장치가 제공될 수 있다.
- [0024] 본 출원의 일 실시예에 따르면, 상기 복수 개의 커널들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고, 상기 커널 식별 모듈은: 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하고, 상기 제 1 레이어로부터 상기 제 1 커널의 위치를 식별하고, 상기 제 1 레이어의 바로 이전 추론 순서인 상기 제 2 레이어로부터 상기 식별된 제 1 커널의 위치와 동일한 위치의 커널을 상기 제 2 커널로 식별하도록 더 설정된, 전자 장치가 제공될 수 있다.
- [0025] 본 출원의 일 실시예에 따르면, 상기 제 1 레이어의 제 1 커널과 상기 제 1 레이어와 바로 이전 추론 순서인 상기 제 2 레이어의 제 1 커널의 위치와 동일한 위치인 제 2 커널과의 가중치의 유사도가, 상기 제 1 커널과 다른 커널과의 가중치의 유사도 보다 큰 전자 장치가 제공될 수 있다.
- [0026] 본 출원의 일 실시예에 따르면, 상기 커널 식별 모듈은, 상기 복수 개의 커널들 중 상기 서로 인접한 상기 제 1 커널 및 상기 제 2 커널 이외의 나머지 커널들 중에서 서로 인접한 커널들을 식별하도록 더 설정되고, 상기 커널 차분 계산 모듈은, 상기 서로 인접한 커널들의 제 1 차분들을 계산하고 저장하도록 더 설정된, 전자 장치가 제공될 수 있다.
- [0027] 본 출원의 일 실시예에 따르면, 상기 계산된 차분은, 상기 제 1 커널 및 상기 제 2 커널에 관련된 인덱스와 함께 저장되고, 상기 계산된 제 1 차분들은, 상기 제 1 차분들의 각각과 관련된 서로 인접한 커널들의 각각과 관련된 인덱스와 함께 저장되는, 전자 장치가 제공될 수 있다.
- [0028] 본 출원의 일 실시예에 따르면, 상기 복수 개의 커널들의 가중치의 분산 값 보다 상기 저장된 계산된 차분과 상기 저장된 계산된 제 1 차분들의 가중치의 분산 값이 더 작은 것을 특징으로 하는, 전자 장치가 제공될 수 있다.
- [0029] 본 출원의 일 실시예에 따르면, 상기 계산된 차분 및 상기 계산된 제 1 차분들은, 양자화 및 지정된 코딩 방식을 기반으로, 저장되는, 전자 장치가 제공될 수 있다.
- [0030] 본 출원의 일 실시예에 따르면, 상기 제 1 레이어의 가중치와 상기 제 2 레이어의 가중치의 차이를 감소시키기 위해 상기 복수 개의 레이어들 간의 손실을 계산하고, 상기 계산된 손실을 목적 함수로 하여, 상기 계산된 손실이 최소화되도록 상기 복수 개의 레이어들에 포함된 커널들의 가중치 값들을 재구성하고, 상기 재구성 이후, 상기 커널들의 가중치 값들의 차분을 저장하도록 설정된, 레이어 손실 계산 모듈;을 더 포함하는, 전자 장치가 제공될 수 있다.
- [0031] 본 출원의 일 실시예에 따르면, 획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델을 저장하는 단계; 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커

널 및 제 2 커널을 식별하는 단계; 및 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하는 단계;를 포함하는, 전자 장치의 제어 방법이 제공될 수 있다.

[0032] 본 출원의 일 실시예에 따르면, 상기 복수 개의 레이어들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고, 상기 제어 방법은 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하는 단계; 상기 복수 개의 레이어들 중 상기 제 1 레이어의 이전의 추론 순서인 제 2 레이어에 포함된 커널들과 상기 제 1 커널 간의 디스턴스를 계산하는 단계; 및 상기 제 2 레이어에 포함된 커널들 중 계산된 디스턴스가 가장 작은 커널을 상기 제 2 커널로 식별하는 단계;를 포함하는, 전자 장치의 제어 방법이 제공될 수 있다.

[0033] 본 출원의 일 실시예에 따르면, 상기 복수 개의 커널들은 상기 입력 커널에 대한 학습을 위해 순차적으로 이용되는 복수 개의 레이어들을 포함하고, 상기 제어 방법은 상기 복수 개의 레이어들 중 제 1 레이어로부터 상기 제 1 커널을 식별하는 단계; 상기 제 1 레이어로부터 상기 제 1 커널의 위치를 식별하는 단계; 및 상기 제 1 레이어의 바로 이전 추론 순서인 상기 제 2 레이어로부터 상기 식별된 제 1 커널의 위치와 동일한 위치의 커널을 상기 제 2 커널로 식별하는 단계;를 포함하는, 전자 장치의 제어 방법이 제공될 수 있다.

[0034] 본 출원의 일 실시예에 따르면, 상기 제 1 레이어의 제 1 커널과 상기 제 1 레이어와 바로 이전 추론 순서인 상기 제 2 레이어의 제 1 커널의 위치와 동일한 위치인 제 2 커널과의 가중치의 유사도가, 상기 제 1 커널과 다른 커널과의 가중치의 유사도 보다 큰, 전자 장치의 제어 방법이 제공될 수 있다.

[0035] 본 출원의 일 실시예에 따르면, 상기 커널 식별 모듈은, 상기 복수 개의 커널들 중 상기 서로 인접한 상기 제 1 커널 및 상기 제 2 커널 이외의 나머지 커널들 중에서 서로 인접한 커널들을 식별하는 단계; 및 상기 서로 인접한 커널들의 제 1 차분들을 계산하고 저장하는 단계;를 더 포함하는 전자 장치의 제어 방법이 제공될 수 있다.

[0036] 본 출원의 일 실시예에 따르면, 상기 계산된 차분은, 상기 제 1 커널 및 상기 제 2 커널에 관련된 인덱스와 함께 저장되고, 상기 계산된 제 1 차분들은, 상기 제 1 차분들의 각각과 관련된 서로 인접한 커널들의 각각과 관련된 인덱스와 함께 저장되는, 전자 장치의 제어 방법이 제공될 수 있다.

[0037] 본 출원의 일 실시예에 따르면, 상기 복수 개의 커널들의 가중치의 분산 값 보다 상기 저장된 계산된 차분과 상기 저장된 계산된 제 1 차분들의 가중치의 분산 값이 더 작은 것을 특징으로 하는, 전자 장치의 제어 방법이 제공될 수 있다.

[0038] 본 출원의 일 실시예에 따르면, 상기 계산된 차분 및 상기 계산된 제 1 차분들은, 양자화 및 지정된 코딩 방식을 기반으로, 저장되는, 전자 장치의 제어 방법이 제공될 수 있다.

[0039] 본 출원의 일 실시예에 따르면, 상기 제 1 레이어의 가중치와 상기 제 2 레이어의 가중치의 차이를 감소시키기 위해 상기 복수 개의 레이어들 간의 손실을 계산하는 동작; 상기 계산된 손실을 목적 함수로 하여, 상기 계산된 손실이 최소화되도록 상기 복수 개의 레이어들에 포함된 커널들의 가중치 값들을 재구성하는 동작; 및 상기 재구성 이후, 상기 커널들의 가중치 값들의 차분을 저장하도록 하는 동작;을 더 포함하는, 전자 장치의 제어 방법이 제공될 수 있다.

[0040] 본 출원의 일 실시예에 따르면, 제어부; 및 획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함하는 심층 신경망 모델 및 인스트럭션들을 저장하는 메모리;를 포함하고, 상기 인스트럭션들이 실행되는 경우, 상기 인스트럭션들은 상기 제어부가: 상기 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별하도록 하고, 상기 제 1 커널의 가중치와 제 2 커널의 가중치의 차분을 계산하고, 상기 계산된 차분을 저장하도록 하는, 전자 장치가 제공될 수 있다.

[0041] 본 명세서에 기재된 실시예는 본 출원이 속하는 기술 분야에서 통상의 지식을 가진 자에게 다양한 실시예들에 따르면, 사상을 명확히 설명하기 위한 것이므로, 본 출원이 본 명세서에 기재된 실시예에 의해 한정되는 것은 아니며, 다양한 실시예들에 따르면, 범위는 다양한 실시예들에 따르면, 사상을 벗어나지 아니하는 수정예 또는 변형예를 포함하는 것으로 해석되어야 한다.

[0042] 본 명세서에서 사용되는 용어는 본 출원에서의 기능을 고려하여 가능한 현재 널리 사용되고 있는 일반적인 용어를 선택하였으나 이는 본 출원이 속하는 기술 분야에서 통상의 지식을 가진 자의 의도, 관례 또는 새로운 기술의 출현 등에 따라 달라질 수 있다. 다만, 이와 달리 특정한 용어를 임의의 의미로 정의하여 사용하는 경우에는 그 용어의 의미에 관하여 별도로 기재할 것이다. 따라서 본 명세서에서 사용되는 용어는 단순한 용어의 명칭이 아닌 그 용어가 가진 실질적인 의미와 본 명세서의 전반에 걸친 내용을 토대로 해석되어야 한다.

[0043] 본 명세서에 첨부된 도면은 본 출원을 용이하게 설명하기 위한 것으로 도면에 도시된 형상은 다양한 실시예들에 따르면, 이해를 돕기 위하여 필요에 따라 과장되어 표시된 것일 수 있으므로 본 출원이 도면에 의해 한정되는 것은 아니다.

[0044] 본 명세서에서 본 출원에 관련된 공지의 구성 또는 기능에 대한 구체적인 설명이 다양한 실시예들에 따르면, 요지를 흐릴 수 있다고 판단되는 경우에 이에 관한 자세한 설명은 필요에 따라 생략하기로 한다.

[0046] 1. 인공 신경망 시스템

[0047] 이하에서는 본 출원의 일 실시예에 따른 인공 신경망 시스템에 대해서 설명한다.

[0048] 본 명세서에서 인공 신경망 시스템은 향상된 압축률을 제공하기 위한, 컨볼루션을 위한 커널의 가중치 차분을 계산하고, 양자화 및 엔트로피 코딩을 기반으로 계산된 가중치 차분을 저장하는 동작을 수행하는 시스템으로 정의될 수 있다.

[0049] 본 명세서에서 인공 신경망 시스템은 향상된 압축률에 기반하여, 소형화된 전자 장치에서 구동될 수 있다.

[0050] 이하에서는 인공 신경망 시스템에 대해서 더 구체적으로 설명한다.

[0052] 2. 인공 신경망 시스템의 구현

[0053] 도 1은 본 출원의 일 실시예에 따른 인공 신경망 시스템이 구현될 수 있는 전자 장치를 설명하기 위한 도면(1)이다.

[0054] 본 출원의 일 실시예에 따른 인공 신경망 시스템은 다양한 종류의 전자 장치들(1)에 구현될 수 있다. 예를 들어, 도 1을 참조하면, 인공 신경망 시스템은 서버 디바이스(100) 및/또는 로컬 디바이스(200)에 구현될 수 있다. 달리 말해, 서버 디바이스(100) 및/또는 로컬 디바이스(200)는 각각의 장치에 구현된 인공 신경망 시스템을 기반으로, 컨볼루션을 위한 커널의 가중치 차분을 계산하고, 양자화 및 엔트로피 코딩을 기반으로 계산된 가중치 차분을 저장하는 동작을 수행할 수 있다. 한편, 본 출원의 일 실시예에 따른 인공 신경망 시스템은, 상기 도 1에 도시된 바에 국한되지 않고, 더 다양한 전자 장치와 서버들에 구현될 수도 있을 것이다.

[0055] 본 출원의 일 실시예에 따른 로컬 디바이스(200)는, 도 1에 도시된 바와 같이, 사용자 단말은 데스크탑 피시, 태블릿 피시, 모바일 단말 등의 개인 사용자에게 의해 이용될 수 있는 장치일 수 있다. 이외에도 유사한 기능을 수행하는 다른 전자 장치들이 로컬 디바이스(200)로 이용될 수 있다.

[0056] 본 출원의 일 실시예에 따른 서버 디바이스(100)는, 복수 개의 로컬 디바이스(200)들과 무선 및/또는 유선 통신을 수행하며, 대단위의 저장 용량을 갖는 데이터베이스를 포함하는 장치일 수 있다. 예를 들어, 서버 디바이스(100)는 복수 개의 로컬 디바이스들과 연동되는 클라우드 디바이스(Cloud device)일 수 있다.

[0057] 상술한 로컬 디바이스(200) 및 서버 디바이스(100)의 구성에 대해서는 구체적으로, 2.2 목차에서 후술한다.

[0059] 2.1. 기능적 구현

[0060] 본 출원의 일 실시예에 따른 인공 신경망 시스템은 동작을 위한 다양한 모듈들을 포함할 수 있다. 인공 신경망 시스템에 포함된 모듈들은 인공 신경망 시스템이 구현되는(또는, 물리적 장치에 포함되는) 물리적 장치(예: 로컬 디바이스(200) 및/또는 서버 디바이스(100))가 지정된 동작을 수행할 수 있도록 구현된 컴퓨터 코드 내지는 인스터션들일 수 있다. 다시 말해, 인공 신경망 시스템이 구현되는 물리적 장치는 복수 개의 모듈들을 컴퓨터 코드 형태로 메모리에 저장하고, 메모리에 저장된 복수 개의 모듈들이 실행되는 경우 복수 개의 모듈들은 물리적 장치가 복수 개의 모듈들에 대응하는 지정된 동작들을 수행하도록 할 수 있다.

[0061] 도 2는 본 출원의 일 실시예에 따른 인공 신경망 시스템에 포함된 모듈들을 설명하기 위한 도면이다.

[0062] 본 출원의 일 실시예에 따른 인공 신경망 시스템은 심층 신경망 모듈(310), 커널 식별 모듈(320), 및 커널 차분 계산 모듈(330)을 포함할 수 있다. 한편, 상기 기재된 바에 국한되지 않고, 인공 신경망 시스템은 기재된 모듈들 이외에 다양한 동작을 수행하도록 구현된 다양한 모듈들을 더 포함하거나, 기재된 모듈들 중 더 적은 모듈들을 포함할 수도 있다. 예를 들어, 인공 신경망 시스템은 inter-layer loss 및 전체 loss를 계산하기 위한 손실 계

산 모듈(미도시)을 더 포함할 수 있다.

- [0063] 본 출원의 일 실시예에 따른 심층 신경망 모듈(310)은 획득되는 입력 데이터에 대한 학습을 위한 복수 개의 레이어들을 포함할 수 있다. 예를 들어, 심층 신경망 모듈(310)은 입력 데이터를 획득하기 위한 적어도 하나의 입력층, 획득된 입력 데이터를 컨볼루션하기 위한 복수 개의 컨볼루션 레이어들, 및 입력 데이터 및 복수 개의 컨볼루션 레이어들에 기반하여 획득되는 출력 데이터를 획득하기 위한 출력층을 포함할 수 있다. 즉, 복수 개의 레이어들은 상술한 복수 개의 컨볼루션 레이어들일 수 있다. 또, 복수 개의 레이어들은, 심층 신경망 모듈의 순차적인 학습을 위한 컨볼루션 레이어를 포함할 수 있는데, 이에 대해서는 3.1 목차의 인공 신경망 시스템의 기본 동작에서 후술한다. 한편, 상기 심층 신경망 모듈(310)은 상기 입력층과 상기 출력층 사이에 위치하는 복수의 은닉층들을 더 포함하며, 상기 은닉층들 사이는 연결 가중치들을 통해 연결되고, 상기 연결 가중치들은 상기 출력 데이터와 상기 입력 데이터에 대응하는 예측 데이터의 차이에 기초하여 갱신되는 은닉층 그룹을 포함하되, 상기 은닉층들 각각은 활성화 함수를 포함하는 은닉 노드를 포함할 수도 있다.
- [0064] 본 출원의 일 실시예에 따른 커널 식별 모듈(320)은 심층 신경망 모듈(310)에 포함된 복수 개의 레이어들 중 차분을 계산하기 위한 커널들을 식별할 수 있다. 심층 신경망 모듈(310)의 커널 식별 동작에 대해서는, 3.1 목차의 인공 신경망 시스템의 기본 동작에서 후술한다.
- [0065] 본 출원의 일 실시예에 따른 커널 차분 계산 모듈(330)은 커널 식별 모듈(320)에 의해 식별된 커널들 간의 차분을 계산하고, 지정된 방식(예: 양자화 및 엔트로피 코딩)을 기반으로 계산된 차분들을 저장할 수 있다. 커널 차분 계산 모듈(330)의 커널 차분 계산 동작에 대해서는, 3.1 목차의 인공 신경망 시스템의 기본 동작에서 후술한다.
- [0066] 한편, 각각의 모듈들은 로컬 디바이스(200) 및/또는 서버 디바이스(100)에 기능적으로 구현될 수 있는데, 이에 대해서는 구체적으로 후술한다.
- [0068] **2.2. 서버 디바이스(100) 및/또는 로컬 디바이스(200)의 구성**
- [0069] 도 3은 본 출원의 일 실시예에 따른 서버 디바이스(100) 및/또는 로컬 디바이스(200)의 구성을 나타내는 블록도이다.
- [0070] 도 3을 참조하면, 서버 디바이스(100) 및/또는 로컬 디바이스(200)는 입/출력부(310), 통신부(320), 데이터베이스(230) 및 제어부(240)를 포함할 수 있다.
- [0071] 로컬입/출력부(310)는 사용자 입력을 받거나 또는 사용자에게 정보를 출력하는 각종 인터페이스나 연결 포트 등일 수 있다. 로컬입/출력부(310)는 입력 모듈과 출력 모듈로 구분될 수 있는데, 입력 모듈은 사용자로부터 사용자 입력을 수신한다. 사용자 입력은 키 입력, 터치 입력, 음성 입력을 비롯한 다양한 형태로 이루어질 수 있다. 이러한 사용자 입력을 받을 수 있는 입력 모듈의 예로는 전통적인 형태의 키패드나 키보드, 마우스는 물론, 사용자의 터치를 감지하는 터치 센서, 음성 신호를 입력받는 마이크, 영상 인식을 통해 제스처 등을 인식하는 카메라, 사용자 접근을 감지하는 조도 센서나 적외선 센서 등으로 구성되는 근접 센서, 가속도 센서나 자이로 센서 등을 통해 사용자 동작을 인식하는 모션 센서 및 그 외의 다양한 형태의 사용자 입력을 감지하거나 입력받는 다양한 형태의 입력 수단을 모두 포함하는 포괄적인 개념이다. 여기서, 터치 센서는 디스플레이 패널에 부착되는 터치 패널이나 터치 필름을 통해 터치를 감지하는 압전식 또는 정전식 터치 센서, 광학적인 방식에 의해 터치를 감지하는 광학식 터치 센서 등으로 구현될 수 있다. 이외에도 입력 모듈은 자체적으로 사용자 입력을 감지하는 장치 대신 사용자 입력을 입력받는 외부의 입력 장치를 연결시키는 입력 인터페이스(USB 포트, PS/2 포트 등)의 형태로 구현될 수도 있다. 또 출력 모듈은 각종 정보를 출력해 사용자에게 이를 제공할 수 있다. 출력 모듈은 영상을 출력하는 디스플레이, 소리를 출력하는 스피커, 진동을 발생시키는 햅틱 장치 및 그 외의 다양한 형태의 출력 수단을 모두 포함하는 포괄적인 개념이다. 이외에도 출력 모듈은 상술한 개별 출력 수단을 연결시키는 포트 타입의 출력 인터페이스의 형태로 구현될 수도 있다.
- [0072] 일 예로, 디스플레이 형태의 출력 모듈은 텍스트, 정지 영상, 동영상을 디스플레이 할 수 있다. 디스플레이는 액정 디스플레이(LCD: Liquid Crystal Display), 발광 다이오드(LED: light emitting diode) 디스플레이, 유기 발광 다이오드(OLED: Organic Light Emitting Diode) 디스플레이, 평판 디스플레이(FPD: Flat Panel Display), 투명 디스플레이(transparent display), 곡면 디스플레이(Curved Display), 플렉시블 디스플레이(flexible display), 3차원 디스플레이(3D display), 홀로그래픽 디스플레이(holographic display), 프로젝터 및 그 외의 영상 출력 기능을 수행할 수 있는 다양한 형태의 장치를 모두 포함하는 광의의 영상 표시 장치를 의미하는 개념

이다. 이러한 디스플레이는 입력 모듈의 터치 센서와 일체로 구성된 터치 디스플레이의 형태일 수도 있다.

[0073] 통신부(320)는 외부 기기와 통신할 수 있다. 따라서, 서버 디바이스(100) 및/또는 로컬 디바이스(200)는 통신부를 통해 외부 기기와 정보를 송수신할 수 있다. 예를 들어, 서버 디바이스(100) 및/또는 로컬 디바이스(200)는 통신부를 이용해 인공 신경망 시스템에 저장되는 계산된 차분 값들이 공유되도록 외부 기기와 통신을 수행할 수 있다.

[0074] 여기서, 통신, 즉 데이터의 송수신은 유선 또는 무선으로 이루어질 수 있다. 이를 위해 통신부는 LAN(Local Area Network)를 통해 인터넷 등에 접속하는 유선 통신 모듈, 이동 통신 기지국을 거쳐 이동 통신 네트워크에 접속하여 데이터를 송수신하는 이동 통신 모듈, 와이파이(Wi-Fi) 같은 WLAN(Wireless Local Area Network) 계열의 통신 방식이나 블루투스(Bluetooth), 직비(Zigbee)와 같은 WPAN(Wireless Personal Area Network) 계열의 통신 방식을 이용하는 근거리 통신 모듈, GPS(Global Positioning System)과 같은 GNSS(Global Navigation Satellite System)을 이용하는 위성 통신 모듈 또는 이들의 조합으로 구성될 수 있다.

[0075] 데이터베이스(330)는 각종 정보를 저장할 수 있다. 데이터베이스는 데이터를 임시적으로 또는 반영구적으로 저장할 수 있다. 예를 들어, 서버(100)의 데이터베이스에는 서버(100)를 구동하기 위한 운영 프로그램(OS: Operating System), 웹 사이트를 호스팅하기 위한 데이터나 점자 생성을 위한 프로그램 내지는 어플리케이션(예를 들어, 웹 어플리케이션)에 관한 데이터 등이 저장될 수 있다. 또, 데이터베이스는 상술한 바와 같이 모듈들을 컴퓨터 코드 형태로 저장할 수 있다. 또, 데이터 베이스는 심층 신경망 모듈(310)에 포함된 복수 개의 레이어들(331)을 저장할 수 있다.

[0076] 데이터베이스(330)의 예로는 하드 디스크(HDD: Hard Disk Drive), SSD(Solid State Drive), 플래쉬 메모리(flash memory), 롬(ROM: Read-Only Memory), 램(RAM: Random Access Memory) 등이 있을 수 있다. 이러한 데이터베이스는 내장 타입 또는 탈부착 가능한 타입으로 제공될 수 있다.

[0077] 제어부(340)는 서버 디바이스(100) 및/또는 로컬 디바이스(200)의 전반적인 동작을 제어한다. 이를 위해 제어부(340)는 각종 정보의 연산 및 처리를 수행하고 서버(100)의 구성요소들의 동작을 제어할 수 있다. 예를 들어, 제어부(340)는 점자 생성을 위한 프로그램 내지 어플리케이션을 실행시킬 수 있을 것이다. 제어부(340)는 하드웨어 소프트웨어 또는 이들의 조합에 따라 컴퓨터나 이와 유사한 장치로 구현될 수 있다. 하드웨어적으로 제어부(340)는 전기적인 신호를 처리하여 제어 기능을 수행하는 전자 회로 형태로 제공될 수 있으며, 소프트웨어적으로는 하드웨어적인 제어부(340)를 구동시키는 프로그램 형태로 제공될 수 있다. 한편, 이하의 설명에서 특별한 언급이 없는 경우에는 서버 디바이스(100) 및/또는 로컬 디바이스(200)의 동작은 제어부(340)의 제어에 의해 수행되는 것으로 해석될 수 있다. 즉, 상술한 인공 신경망 시스템에 구현되는 모듈들이 실행되는 경우, 모듈들은 제어부(340)가 서버 디바이스(100) 및/또는 로컬 디바이스(200)를 이하의 동작들을 수행하도록 제어하는 것으로 해석될 수 있다.

[0079] 2.3. 인공 신경망 시스템의 구현예

[0080] 인공 신경망 시스템은 네트워크 타입 또는 온 디바이스(On-device) 타입으로 구현될 수 있다.

[0081] 예를 들어, 상술한 모듈들이 복수 개의 물리적 장치들에 구현되는 경우, 인공 신경망 시스템은 네트워크 타입으로 정의될 수 있다. 일 예로, 심층 신경망 모듈(310)이 로컬 디바이스(200)에 구현되고, 연산을 위한 커널 식별 모듈(320) 및 커널 차분 계산 모듈(330)이 서버 디바이스(100)에 구현되는 경우, 인공 신경망 시스템은 네트워크 타입으로 정의될 수 있다.

[0082] 예를 들어, 상술한 모듈들이 하나의 물리적 장치에 구현되는 경우, 인공 신경망 시스템은 온 디바이스(On-device) 타입으로 정의될 수 있다. 일 예로, 심층 신경망 모듈(310), 커널 식별 모듈(320) 및 커널 차분 계산 모듈(330)이 로컬 디바이스(200)에만 구현될 수 있다.

[0083] 다만, 본 출원의 일 실시예에 따른 인공 신경망 시스템은 보다 향상된 압축률을 제공할 수 있으므로, 보다 안정적으로 온-디바이스(On-device) 타입으로 구동될 수 있다.

[0085] 3. 인공 신경망 시스템의 동작

[0086] 3.1. 커널간 차분 저장 동작(INTER-LAYER WEIGH PREDICTION)

- [0087] 도 4는, 본 출원의 일 실시예에 따른 인공 신경망 시스템의 커널간 차분 저장 동작의 일 예를 설명하기 위한 흐름도(400)이다. 다양한 실시예들에 따르면 인공 신경망 시스템의 동작은 도 4에 도시되는 동작의 순서에 국한되지 않고, 도시되는 순서와 다른 순서로 수행될 수 있다. 또한, 다양한 실시예들에 따르면, 도 4에 도시되는 인공 신경망 시스템의 동작 보다 더 많은 동작들이 수행되거나, 또는 더 적은 적어도 하나의 동작이 수행될 수도 있다. 이하에서는 도 5 내지 8을 참조하여 도 4에 대해서 설명한다.
- [0088] 도 5는 본 출원의 일 실시예에 따른 인공 신경망 시스템의 커널간 차분 저장 동작을 설명하기 위한 도면이다.
- [0089] 먼저, 410 동작에 대해서 설명한다.
- [0090] 본 출원의 일 실시예에 따르면, 인공 신경망 시스템(예: 커널 식별 모듈(320))은 410 동작에서 심층 신경망 모델에 포함된 복수 개의 레이어들을 식별할 수 있다.
- [0091] 인공 신경망 모델에 포함된 복수 개의 레이어들은 도 5에 도시된 바와 같이 컨볼루션을 위한 가중치를 갖는 컨볼루션 레이어들일 수 있다. 복수 개의 레이어들은 각각이 특정 인공 신경망 모델에서, 획득된 입력 데이터에 대한 추론을 위해 서로 다른 순서로 이용될 수 있다. 구체적으로, 상기 서로 다른 레이어는 도 5에 도시된 바와 같이 학습이 완료된 인공 신경망 모델에서 추론을 위한 제 1 컨볼루션 레이어(311)(예: uth layer) 및 제 2 컨볼루션 레이어(313)(예: ith layer, $u < i$)를 포함하고, 제 2 컨볼루션 레이어(313)의 추론 순서는 제 1 컨볼루션 레이어(311)의 추론 순서 보다 빠를 수 있다. 복수 개의 레이어들 각각(예: 311, 313)은 컨볼루션을 위한 복수 개의 커널들(예: 314, 312)을 포함할 수 있다.
- [0092] 본 출원의 일 실시예에 따르면, 상기 컨볼루션 레이어들의 컨볼루션 방식은 grouped convolution, depthwise convolution, pointwise convolution, depthwise separable convolution 등의 다양한 컨볼루션을 포함할 수 있다. 한편, 본 출원의 인공 신경망 시스템의 동작은 depthwise separable convolution에 더 효과적으로 적용될 수 있는데, 이와 관련된 동작은 3.2 목차에서 더 후술한다.
- [0093] 이하에서는, 420 동작에 대해서 설명한다.
- [0094] 본 출원의 일 실시예에 따르면, 인공 신경망 시스템(예: 커널 식별 모듈(320))은 420동작에서 복수 개의 레이어들 중 각각의 서로 다른 레이어로부터 가중치가 유사한 제 1 커널 및 제 2 커널을 식별할 수 있다. 예를 들어, 커널 식별 모듈(320)은, 도 5에 도시된 바와 같이 인공 신경망 모델에 포함된 복수 개의 레이어들 중에서 서로 다른 레이어(311, 313)로부터 두 커널(312, 314)을 식별할 수 있다. 구체적으로, 커널 식별 모듈(320)은 도 5에 도시된 바와 같이, 제 1 컨볼루션 레이어(311)로부터 제 1 커널(312)을 식별하고, 이미 학습된(또는, 제 1 컨볼루션 레이어(311) 보다 추론 순서가 빠른) 제 2 컨볼루션 레이어(313)로부터 제 2 커널(314)을 식별할 수 있다. 이때, 커널 식별 모듈(320)은, 이미 학습된 컨볼루션 레이어로부터 상기 선택된 커널과 가중치가 가장 유사한(달리 말하면, distance가 가장 작은) 커널을 식별할 수 있다.
- [0095] 구체적으로 본 출원의 일 실시예에 따르면, 커널 식별 모듈(320)은 상술한 가중치가 가장 유사한 커널로서 특정 컨볼루션 레이어에 바로 인접한 컨볼루션 레이어로부터 커널을 식별할 수 있다. 이는, 도 6 및 도 7을 참조하여 설명한다.
- [0096] 도 6은 본 출원의 일 실시예에 따른 인공 신경망 시스템의 가중치 차분 계산을 위한 레이어를 식별하는 동작을 구체적으로 설명하기 위한 도면이다. 도 7은 본 출원의 일 실시예에 따른 인접 레이어의 가중치들의 관계를 설명하기 위한 도면이다.
- [0097] 본 출원의 일 실시예에 따르면, 커널 식별 모듈(320)은 도 6에 도시된 바와 같이, 특정 컨볼루션 레이어의 특정 커널을 식별하고, 특정 컨볼루션 레이어의 바로 이전 추론 순서의 컨볼루션 레이어로부터 특정 커널과 같은 위치의 커널을 가중치 차분 계산을 위한 커널로서 식별할 수 있다. 예를 들어, 커널 식별 모듈(320)은 도 6에 도시된 바와 같이 $i+1$ 번째 레이어로부터 제 1 커널(312)을 식별하고, 특정 커널의 위치 정보를 식별할 수 있다. 커널 식별 모듈(320)은 $i+1$ 번째 레이어의 바로 이전 순서인 i 번째 레이어를 식별하고, i 번째 레이어로부터 특정 커널의 위치 정보에 대응하는(즉, 위치가 동일한) 제 2 커널을 식별할 수 있다.
- [0098] 위와 같은, 커널 식별 모듈(320)의 동작은 도 7에 개시되는 레이어들 간의 가중치의 관계로부터 확인되는 인접 레이어의 가중치들이 유사하고 가중치가 레이어의 순서에 따라 서서히 변한다는 가설(SMOOTHLY VARYING WEIGHT HYPOTHESIS)에 기반할 수 있다.
- [0099] 도 7을 참조하면, 도 7 (a) 내지 (d)는 CIFAR 10 또는 CIFAR 100에 의해 학습된 인공 신경망 모델(MobileNet)에 포함된 소스 레이어(Source layer) 및 타겟 레이어(Target layer) 간의 예측도를 나타내는 히트맵을 개시한

다. 구체적으로, 도 7 (a)는 CIFAR-10 에 의해 학습된 제 1 인공 신경망 모델(MobileNet)의 레이어간 예측도를 나타내고, 도 7 (b)는 CIFAR-10 에 의해 학습된 제 2 인공 신경망 모델(MobileNet V2)의 레이어간 예측도를 나타내고, 도 7 (c)는 CIFAR-100에 의해 학습된 제 1 인공 신경망 모델(MobileNet)의 레이어간 예측도를 나타내고, 도 7 (d)는 CIFAR-100 에 의해 학습된 제 2 인공 신경망 모델(MobileNet V2)의 레이어간 예측도를 나타낸다.

[0100] 도 7 (a) 내지 (d)를 참조하면, 소스 레이어들 중 i th layer와 타겟 레이어들 중 $i-1$ th layer 간의 distance가 가장 짧은 경향이 개시된다. 즉, 이로부터 특정 레이어의 특정 커널과 distance가 가장 짧은 커널은 바로 이전 레이어에 존재하고, 컨볼루션 커널들의 가중치는 레이어를 따라서 부드럽게 변화한다는 것이 도출될 수 있다.

[0101] 이에 따라, 상술한 커널 식별 모듈(320)의 특정 레이어의 바로 이전 순서의 레이어를 식별하는 동작은, 모델의 최적의 압축 효율을 발생시킬 수 있다. 또한, 커널 간의 가중치의 유사도를 계산하는 번거로운 동작의 수행 없이, 커널 식별 모듈(320)은 특정 레이어의 바로 이전 순서의 레이어로부터 특정 커널의 위치에 대응하는 커널을 식별하는 동작을 수행함으로써, 전자 장치의 운용 부담이 경감되는 효과가 발생될 수 있다.

[0102] 한편, 도 7에 의한 결과물은 아래와 같은 [수학식 1]에 의해 표현될 수 있다.

수학식 1

$$P[L(K_{i,j}; K_{i-1,v}) > L(K_{i,j}; K_{u,v})] > P[L(K_{i,j}; K_{i-1,v}) \leq L(K_{i,j}; K_{u,v})], 1 \leq u < i-1$$

[0103]

[수학식 1]에서, K 는 커널, $P[-]$ 는 확률, $L(-,-)$ 는 두 커널 간의 distance, 및 (u,v) 이전 추론 순서의 레이어와 커널의 지수를 나타낼 수 있다.

[0104]

[수학식 1]로부터, 상술한 바와 같이 특정 레이어의 특정 커널과 distance가 가장 짧은 커널은 바로 이전 레이어에 존재하고, 컨볼루션 커널들의 가중치는 레이어를 따라서 부드럽게 변화한다는 것이 도출되므로, 중복되는 설명은 생략한다.

[0105]

이하에서는, 430 동작에 대해서 설명한다.

[0106]

본 출원의 일 실시예에 따른 인공 신경망 시스템(예: 커널 차분 계산 모듈(330))은 430 동작에서 제 2 커널의 가중치와 제 1 커널의 가중치의 차분을 계산하고, 계산된 가중치의 차분을 저장할 수 있다. 예를 들어, 커널 차분 계산 모듈(330)은 도 5에 도시된 바와 같이 상술한 바와 같이 식별된 제 2 커널과 제 1 커널의 가중치의 차분(510)을 계산하고, 데이터 베이스에 저장할 수 있다.

[0107]

본 출원의 일 실시예에 따르면, 커널 차분 계산 모듈(330)은 계산된 가중치의 차분(510)을 소정의 지수와 연관된 형태로 저장할 수 있다. 예를 들어, 커널 차분 계산 모듈(330)은 커널 차분(510)을 레이어의 순서 및 레이어에서 선택된 커널의 순서와 연관된 형태로 저장할 수 있다. 예를 들어, 커널 차분(510)은 도 5의 520과 같이, 이전 레이어(예: 제 2 레이어(313))를 나타내는 순서(예: u) 및 이전 레이어에서 선택된 커널(예: 제 2 커널(314))의 순서(예: v)와 연관된 형태로 저장될 수 있다. 이에 따라, 인공 신경망 시스템은, 특정 레이어의 커널, 특정 레이어의 커널과 관련된 가중치의 차분(510) 및 상기 지수(u, v)를 기반으로, 이전 추론 순서의 레이어의 커널을 획득할 수 있다.

[0108]

본 출원의 일 실시예에 따르면, 커널 차분 계산 모듈(330)은 계산된 가중치를 지정된 방식(예: 양자화 및 엔트로피 코딩 방식)을 기반으로 처리하여 저장할 수 있는데, 이에 대해서는 3.2 목차에서 구체적으로 후술한다.eeee

[0109]

한편, 상술한 특정 컨볼루션 레이어로부터 특정 커널에 대한 가장 인접한 커널을 식별하고 가중치의 차분을 계산하고 저장하는 동작은, 특정 커널에 대해서만 수행되는 것이 아니다. 즉, 특정 컨볼루션 레이어에 포함된 특정 커널 이외의 다른 모든 커널들에 대해서 상술한 동작이 수행될 수 있고, 특정 컨볼루션 레이어가 아닌 다른 미학습된 컨볼루션 레이어에 대해서도 상술한 동작이 수행될 수 있다.

[0110]

3.1.1. 커널간 차분 저장 동작의 효과

[0112]

- [0113] 상술한 커널간 차분 저장 동작에 의해 저장된 데이터의 분산이 감소되고, 이에 따라 압축률이 향상될 수 있다
- [0114] 도 8은 가중치의 분포 및 가중치의 차분의 분포를 나타내는 그래프이다. 구체적으로 도 8 (a)가 가중치의 분포를 나타내고, 도 8 (b)가 가중치의 차분의 분포를 나타낸다.
- [0115] 도 8 (a)에 개시되는 단순히 데이터 베이스에 저장된 가중치의 분포와 도 8 (b)에 개시되는 데이터 베이스에 저장된 가중치 차분의 분포를 비교하면, 가중치 차분의 분포가 더 조밀하며 이에 따라 상대적으로 저장된 데이터의 분산이 감소되는 것이 개시된다.
- [0116] 이에 따라, 상술한 커널간 차분 저장 동작에 의해 모델의 압축률이 향상될 수 있다.

[0118] 3.2. 인접 레이어 손실 (INTER-LAYER LOSS) 계산

- [0119] 본 출원의 일 실시예에 따른 인공 신경망 시스템은, 압축 효율을 더욱 높이기 위해 인접 레이어의 컨볼루션 간의 차분을 줄이기 위하여, inter-layer loss를 기반으로 학습을 수행할 수 있다. inter-layer loss는 첫 번째 레이어를 제외한 각 컨볼루션 레이어의 한 커널이 있을 때, 바로 이전 레이어 동일 위치의 커널과의 distance를 줄여주는 loss일 수 있다. inter-layer loss는 심층신경망 모델 학습 시, classification loss와 함께 작동될 수 있다.
- [0120] 이때, 인공 신경망 시스템은 inter-layer loss 및 전체 loss를 계산하기 위한 손실 계산 모듈(미도시)을 더 포함할 수 있다.
- [0121] 도 9는 본 출원의 일 실시예에 따른 적어도 하나의 목표 함수 (예: inter-layer loss)를 기반으로 학습을 수행하는 인공 신경망 시스템의 동작의 일 예를 설명하기 위한 흐름도(900)이다. 다양한 실시예들에 따르면 인공 신경망 시스템의 동작은 도 9에 도시되는 동작의 순서에 국한되지 않고, 도시되는 순서와 다른 순서로 수행될 수 있다. 또한, 다양한 실시예들에 따르면, 도 9에 도시되는 인공 신경망 시스템의 동작 보다 더 많은 동작들이 수행되거나, 또는 더 적은 적어도 하나의 동작이 수행될 수도 있다. 이하에서는 도 10 내지 11을 참조하여 도 9에 대해서 설명한다.
- [0122] 본 출원의 일 실시예에 따르면, 인공 신경망 시스템은 910동작에서 더 향상된 압축률을 위해, 특정 심층 신경망 모델의 훈련을 위해 목적 함수(objective function)로서 inter-layer loss를 식별할 수 있다. 인공 신경망 시스템은, 상기 기제에 국한되지 않고, 상기 inter-layer loss를 loss function 또는 cost function으로 식별할 수 있다.
- [0123] 예를 들어, 인공 신경망 시스템은, depth-wise separable convolutions의 spatial convolution에서 inter-layer loss를 계산할 수 있다. 구체적으로, 인공 신경망 시스템은 아래와 같은 [수학식 2]를 기반으로 inter-layer loss($L_{\text{inter-layer}}$)를 계산할 수 있다.

수학식 2

$$L_{\text{inter-layer}} = \frac{1}{Z} \sum_{i=2}^N \sum_{j=0}^{c_i-1} |K_{i,j} - K_{i-1,v}|$$

- [0124]
- [0125] 여기서, Z는 예측되는 가중치의 수이고, N은 depth-wise separable convolutions 레이어의 수이고, v는 이전 레이어의 인덱스를 나타낸다.
- [0126] 상기과 같이 계산된 inter-layer loss는 첫 번째 레이어를 제외한 각 컨볼루션 레이어의 한 커널이 있을 때, 바로 이전 레이어 동일 위치의 커널과의 distance를 감소시킬 수 있다. 이에 따라, 인공 신경망 시스템은, 이전 레이어의 커널의 가중치로부터 다음 현재 레이어의 가중치를 예측할 수 있기 때문에 최상의 예측의 지표에 대한 모든 비-텍스처 비트를 제거할 수 있다.
- [0127] 또 예를 들어, 인공 신경망 시스템은 인공 신경망 학습을 위해, 아래와 같이 [수학식 3]을 이용하여, 전체

$\text{loss}(L_{\text{total}})$ 를 계산할 수 있다.

수학식 3

$$L_{\text{total}} = L_{\text{cls}} + \lambda L_{\text{inter-layer}};$$

여기서, L_{cls} 는 He et al., 2016에 개시되는 종래의 크로스 엔트로피 로스 방식을 이용하여 계산된 loss이고, λ 는 계산된 inter-layer loss의 값을 제어하기 위한 파라미터이다. 이때 구체적으로, λ 는 1로 설정될 수 있다. 이에 대해서는 도 10을 참조하여 설명한다.

도 10은 λ (목적함수에서의 inter-layer loss에 대한 가중치)에 따른 인공 신경망 모델 파라미터의 데이터의 양 및 학습 정확도를 나타내는 그래프이다.

도 10 (a)에서 8-bit 그래프는 양자화 및 허프만 코딩 이후 학습된 인공 신경망 모델의 λ 대비 저장된 파라미터(예: 가중치의 차분)의 데이터 양을 나타내고, 도 10 (b)에서 8-bit 그래프는 양자화 및 허프만 코딩 이후 학습된 인공 신경망 모델의 λ 대비 test 정확도를 나타낸다.

이때, 도 10 (b)를 참조하면 λ 가 1인 경우 상대적으로 test 정확도가 적을 수 있으나, 도 10 (a)를 참조하면 λ 가 1인 경우 상대적으로 저장된 파라미터(예: 가중치의 차분)의 데이터 양이 가장 적을 수 있다. 이에 따라, 인공 신경망 모델의 데이터 압축률 관점에서, λ 가 1인 경우 효과가 최대일 수 있다.

한편, 도 10 (a)에 개시되는 inter-layer loss가 적용되지 않은 모델의 파라미터의 데이터의 양을 나타내는 Base line 그래프와 상술한 8-bit 그래프를 비교하면, inter-layer loss가 적용된 모델의 경우가 데이터 압축률 관점에서 더 효과가 있음을 알 수 있다.

본 출원의 일 실시예에 따르면, 인공 신경망 시스템은 920동작에서 목적 함수인 inter-layer loss가 최소화되도록, 심층 신경망 모델에 포함된 레이어의 커널들의 가중치 값들을 재구성할 수 있다. 즉, 인공 신경망 시스템은 인공 신경망 모델에 포함된 레이어의 커널들의 가중치 값들 간의 차이가 최소화되도록 할 수 있다.

예를 들어, 인공 신경망 시스템은, 아래와 같은 [수학식 4]에 기재된 바와 같이, 이전 레이어의 가중치(예: 제 2 컨볼루션 레이어(313))($\tilde{R}_{i-1,j}$) 및 지정된 코딩 방식 및 양자화에 기반하여 저장된 이전 레이어와 다음 레이어의 가중치의 차분($\tilde{R}_{i,j}$)을 기반으로, 다음 순서의 레이어(예: 제 1 컨볼루션 레이어(311))의 가중치를 재구성할 수 있다. 상기 지정된 코딩 방식은 허프만코딩, 산술 부호화, LZW 부호화 등의 다양한 엔트로피 코딩 방식을 포함하며, 바람직하게 허프만 코딩 방식이 이용될 수 있다.

수학식 4

$$\hat{R}_{i,j} = \tilde{R}_{i,j} + \hat{R}_{i-1,j}$$

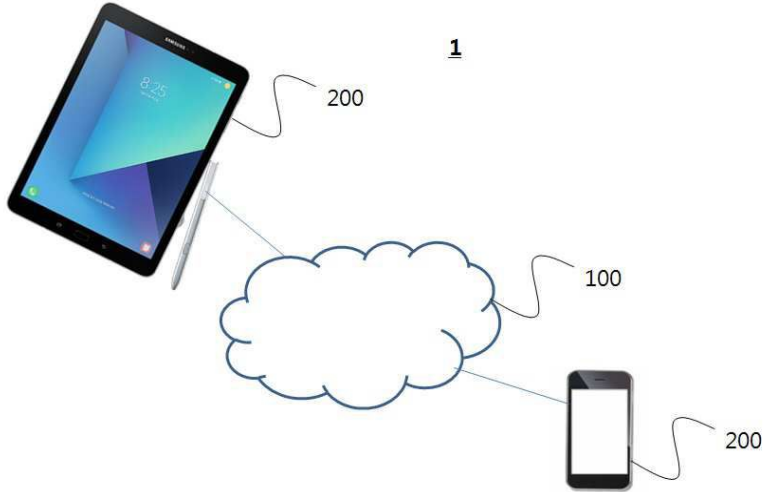
본 출원의 일 실시예에 따르면, 인공 신경망 시스템은 930동작에서 재구성 이후, 심층 신경망 모델의 레이어들의 각각의 커널들의 가중치 값들의 차분을 계산 및 저장할 수 있다. 인공 신경망 시스템의 930 동작은, 상술한 인공 신경망 시스템의 410 동작 내지 430 동작의 커널 간의 가중치 차분을 계산하고 학습하는 동작과 같이 수행될 수 있으므로, 중복되는 설명은 생략한다. 다만, 930 동작에서 저장된 가중치 값들의 차분의 데이터 양은, 430 동작에서 저장된 가중치 값들의 차분의 데이터 양 보다 작을 수 있다. 달리 말해, 상술한 바와 같은 inter-layer loss를 목적 함수로 한 학습에 따라, 인공 신경망 모델의 저장 효율(또는 압축률)이 향상될 수 있다.

- [0138] 도 11은 본 출원의 일 실시예에 따른 인접 레이어의 가중치들의 관계를 설명하기 위한 도면이다.
- [0139] 도 11을 참조하면, 도 11 (a) 내지 (d)는 CIFAR 10 또는 CIFAR 100에 의해 학습된 상술한 손실 계산에 따른 재구성된 커널들을 포함하는 인공 신경망 모델(MobileNet)에 포함된 소스 레이어(Source layer) 및 타겟 레이어(Target layer) 간의 예측도를 나타내는 히트맵을 개시한다. 구체적으로, 도 11 (a)는 CIFAR-10에 의해 학습된 재구성된 커널들을 포함하는 제 1 인공 신경망 모델(MobileNet)의 레이어간 예측도를 나타내고, 도 11 (b)는 CIFAR-10에 의해 학습된 재구성된 커널들을 포함하는 제 2 인공 신경망 모델(MobileNet V2)의 레이어간 예측도를 나타내고, 도 11 (c)는 CIFAR-100에 의해 학습된 재구성된 커널들을 포함하는 제 1 인공 신경망 모델(MobileNet)의 레이어간 예측도를 나타내고, 도 11 (d)는 CIFAR-100에 의해 학습된 재구성된 커널들을 포함하는 제 2 인공 신경망 모델(MobileNet V2)의 레이어간 예측도를 나타낸다.
- [0140] 도 7의 개시 사항과 도 11의 개시 사항을 비교하면, 재구성된 커널들을 포함하는 인공 신경망 모델에서 특정 레이어의 특정 커널과 distance가 가장 짧은 커널은 바로 이전 레이어에 존재하고, 컨볼루션 커널들의 가중치는 레이어를 따라서 부드럽게 변화하는 효과가 더 강화될 수 있다. 이에 따라, 모델의 최적의 압축 효율이 증가하고 및 전자 장치의 운용 부담이 경감되는 효과가 더 강화될 수 있다.
- [0142] **3.2.1. 인접 레이어 손실 (INTER-LAYER LOSS) 계산의 효과**
- [0143] 상술한 저장된 데이터의 더 분산이 감소되고, 이에 따라 압축률이 향상될 수 있다
- [0144] 도 12는 심층신경망 학습 시 inter-layer loss가 목적함수에 포함되지 않은 모델과 심층신경망 학습 시 inter-layer loss가 목적함수에 포함된 모델의 가중치에 대한 차분의 분포를 나타내는 그래프이다. 구체적으로 도 12 (a)가 inter-layer loss가 계산되지 않은 모델의 가중치의 차분의 분포를 나타내고, 도 12 (b)가 inter-layer loss가 계산된 모델의 가중치의 차분의 분포를 나타낸다.
- [0145] 도 12 (a)에 개시되는 단순히 데이터 베이스에 저장된 가중치의 차분의 분포와 도 12 (b)에 개시되는 데이터 베이스에 저장된 가중치 차분의 분포를 비교하면, inter-layer loss가 계산된 모델의 경우가 가중치 차분의 분포가 더 조밀하며 이에 따라 상대적으로 저장된 데이터의 분산이 감소되는 것이 개시된다.
- [0146] 이에 따라, 상술한 inter-layer loss가 계산된 모델의 압축률이 더 향상될 수 있다.
- [0148] **4. 인공 신경망 시스템의 동작에 따른 효과 비교**
- [0149] 이하에서는, 도 13 내지 도 14를 참조하여, 상술한 인공 신경망 시스템의 동작의 효과를 비교한다.
- [0150] 도 13은 본 출원의 일 실시예에 따른 다양한 인공 신경망 모델들의 파라미터의 데이터 양을 나타내는 도면이다. 여기서, ILWP-FSS는 full search strategy에 기반한 학습 모델, ILWP-LSS는 local search strategy에 기반한 학습 모델, ILWP-ILL는 inter-layer loss에 기반하나 search를 수행하지 않는 학습 모델을 나타낸다.
- [0151] 구체적으로, 도 13 (a)는 CIFAR-10에 의해 학습된 제 1 인공 신경망 모델(MobileNet)의 파라미터의 데이터 양을 나타내고, 도 13 (b)는 CIFAR-10에 의해 학습된 제 2 인공 신경망 모델(MobileNet V2)의 파라미터의 데이터 양을 나타내고, 도 13 (c)는 CIFAR-100에 의해 학습된 제 1 인공 신경망 모델(MobileNet)의 파라미터의 데이터 양을 나타내고, 도 13 (d)는 CIFAR-100에 의해 학습된 제 2 인공 신경망 모델(MobileNet V2)의 파라미터의 데이터 양을 나타낸다.
- [0152] 도 13을 참조하면, 목차 3.2에 기반한 inter-layer loss를 이용하는 ILWP-ILL 모델의 파라미터의 데이터 양이 가장 적은 것을 알 수 있다.
- [0153] 도 14는 본 출원의 일 실시예에 따른 다양한 인공 신경망 모델들의 파라미터의 데이터 양 대비 테스트 정확도를 나타내는 도면이다. 여기서, ILWP-FSS는 full search strategy에 기반한 학습 모델, ILWP-LSS는 local search strategy에 기반한 학습 모델, ILWP-ILL는 inter-layer loss에 기반하나 search를 수행하지 않는 학습 모델을 나타낸다.
- [0154] 도 14 (a) 내지 (d)는 상술한 도 13 (a) 내지 (d)의 학습 모델들의 파라미터의 데이터 양 대비 테스트 정확도를 나타내므로, 중복되는 설명은 생략한다.

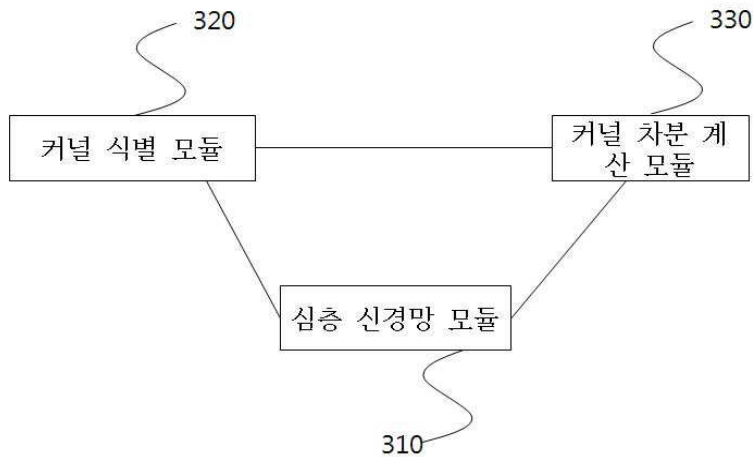
[0155] 도 14를 참조하면, 목차 3.2에 기반한 inter-layer loss를 이용하는 ILWP-ILL 모델이 파라미터의 데이터 양도 적으며, 비교적 양호한 테스트 정확도를 가짐을 알 수 있다.

도면

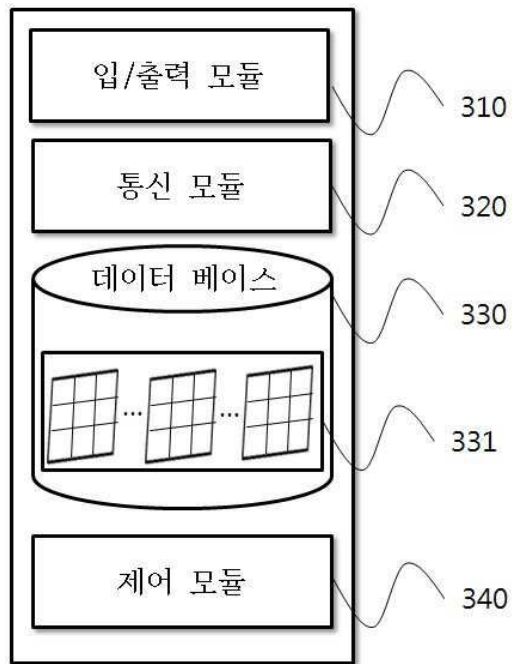
도면1



도면2

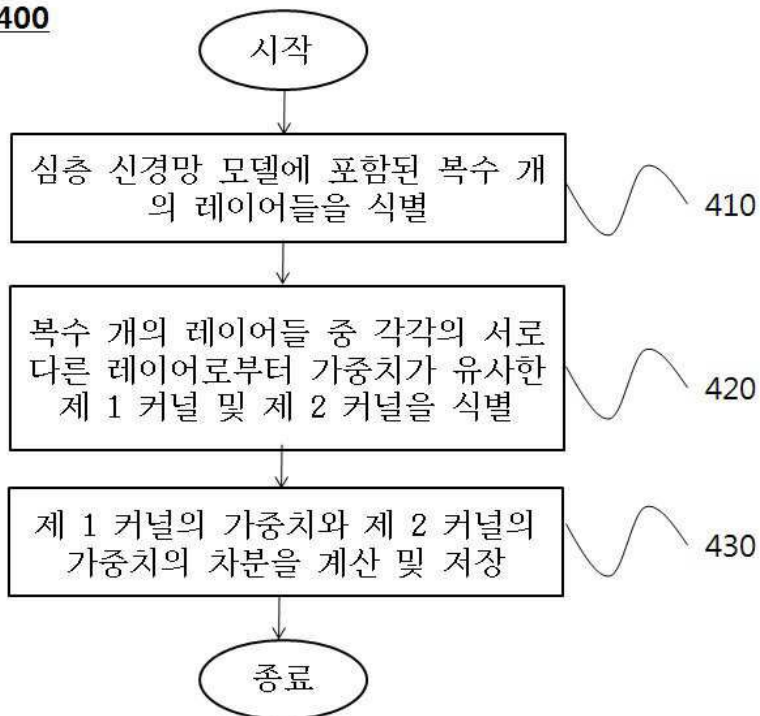


도면3

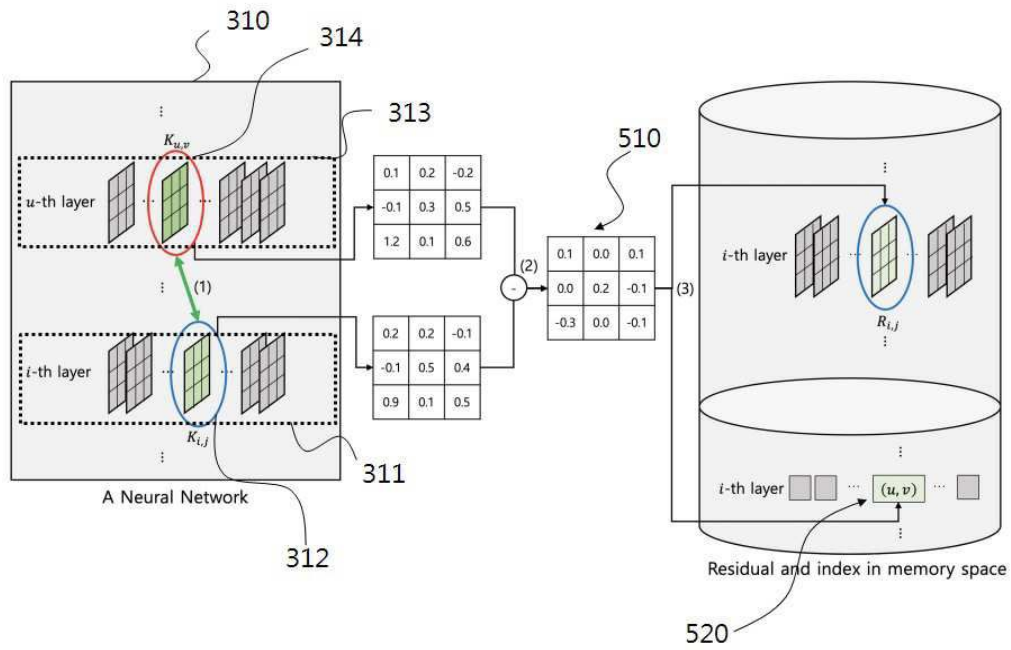


도면4

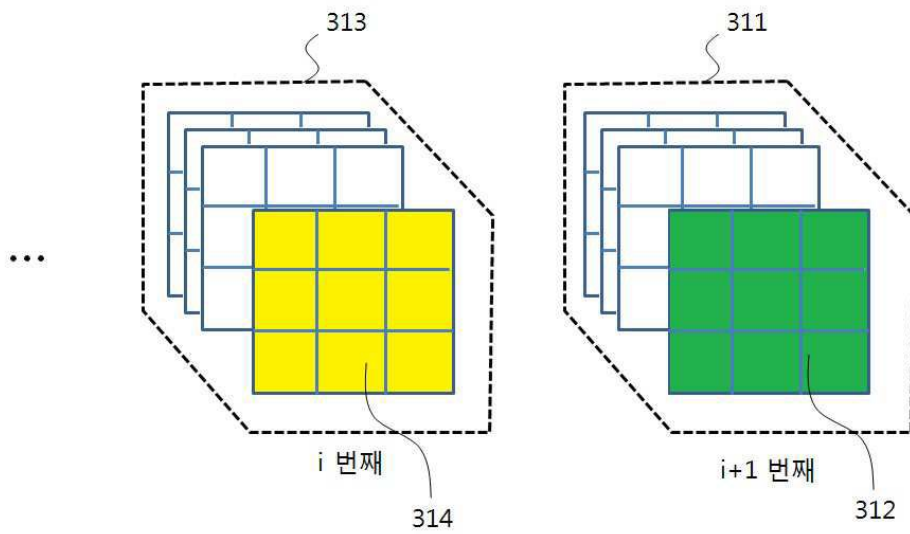
400



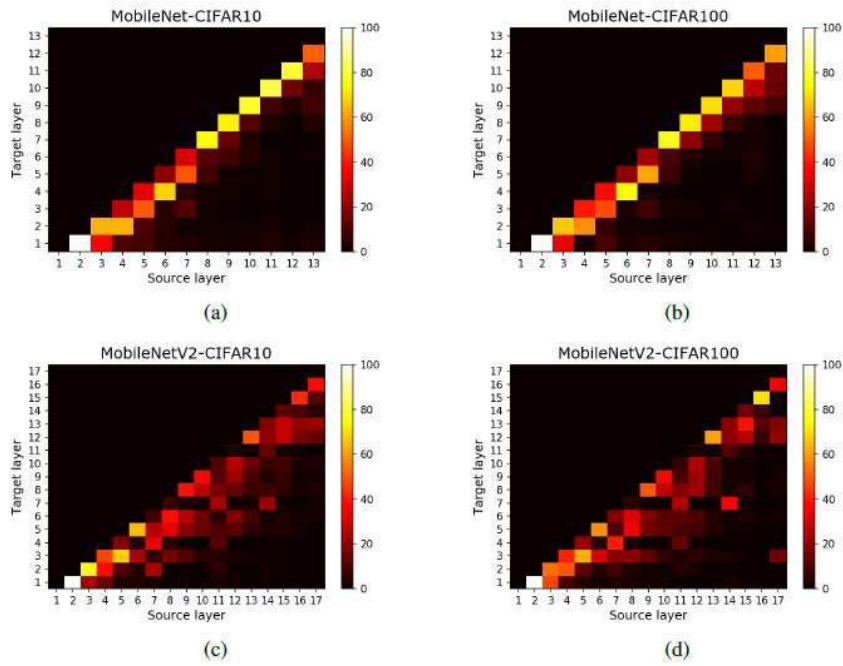
도면5



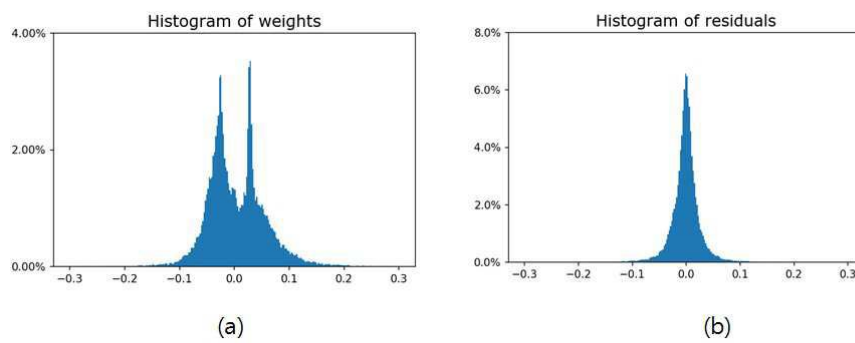
도면6



도면7

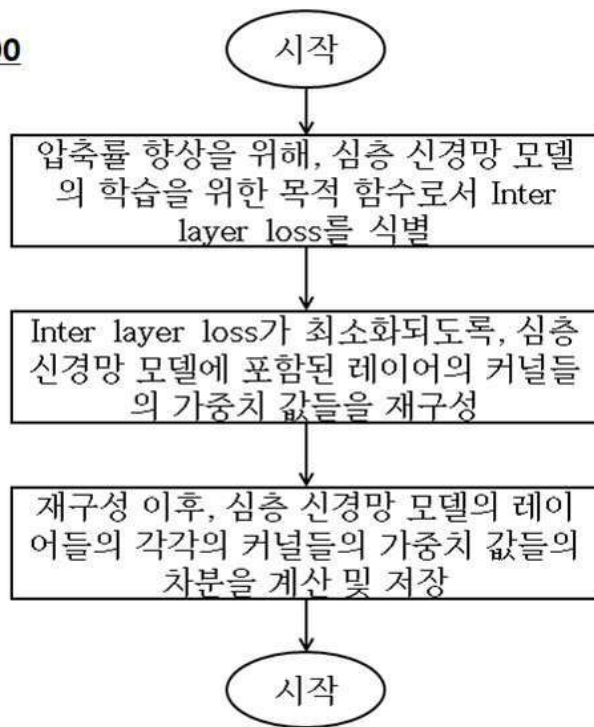


도면8

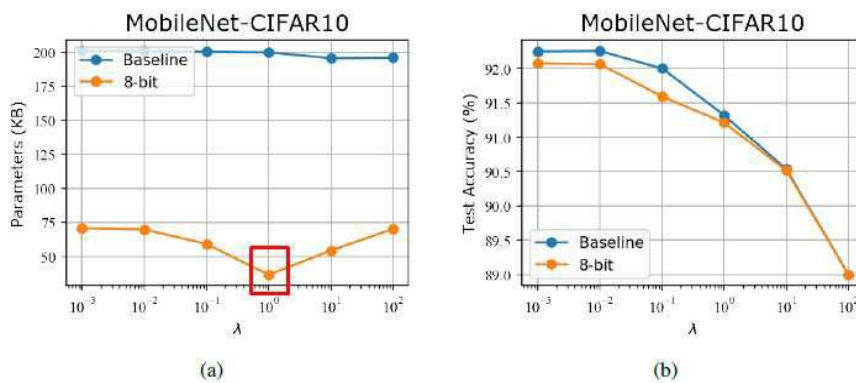


도면9

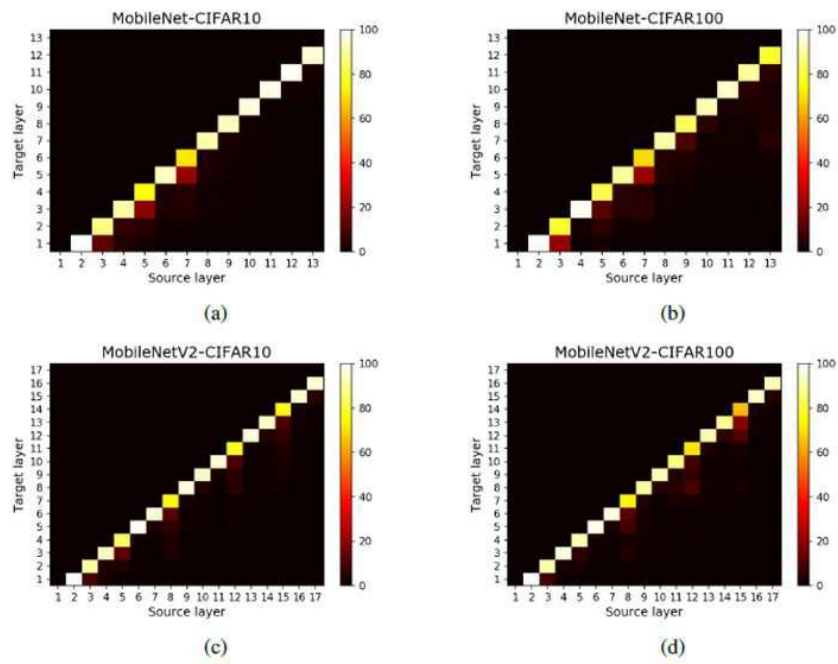
900



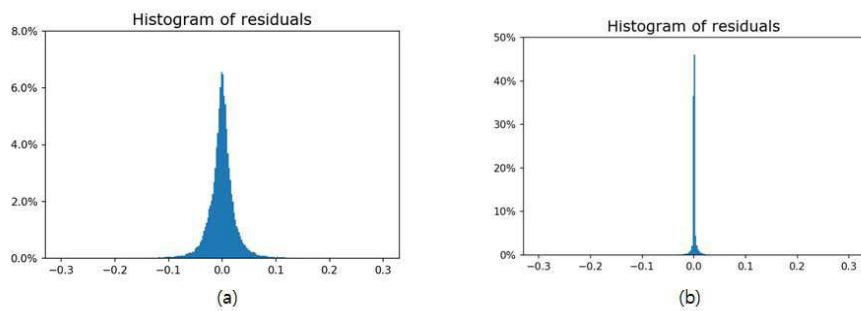
도면10



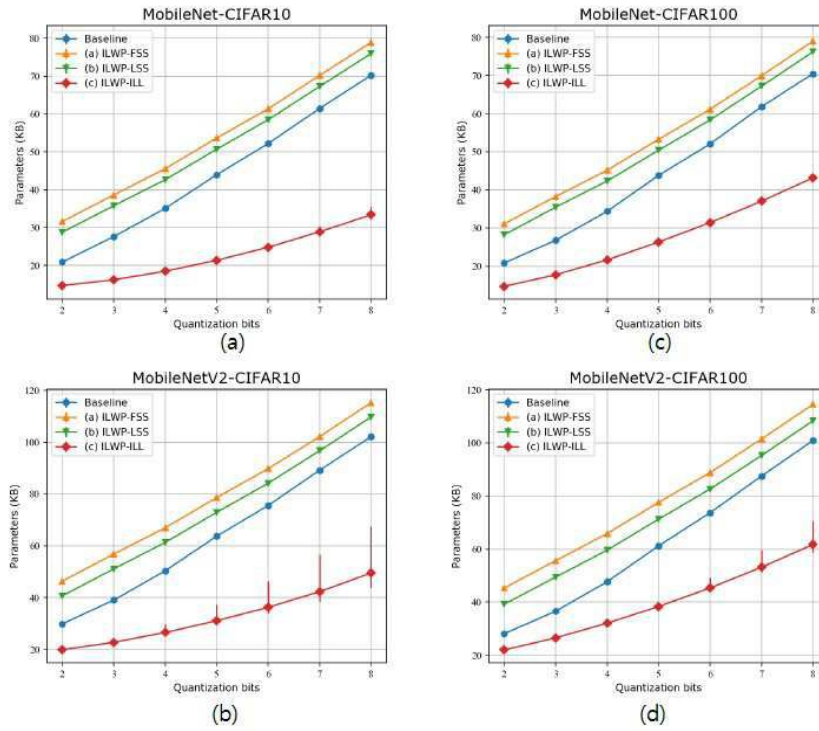
도면11



도면12



도면13



도면14

