



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0127224
(43) 공개일자 2024년08월22일

(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) G06N 3/08 (2023.01)
(52) CPC특허분류
G06N 3/0495 (2023.01)
G06N 3/082 (2023.01)
(21) 출원번호 10-2023-0032854
(22) 출원일자 2023년03월13일
심사청구일자 2023년03월13일
(30) 우선권주장
1020230020347 2023년02월15일 대한민국(KR)

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
박형곤
경기도 고양시 일산동구 일산로 241, 110동 204호(마두동, 백마마을1단지아파트)
이주홍
서울특별시 은평구 가좌로9길 19-1, 401호(응암동, 다베로)
(74) 대리인
특허법인(유한)아이시스

전체 청구항 수 : 총 9 항

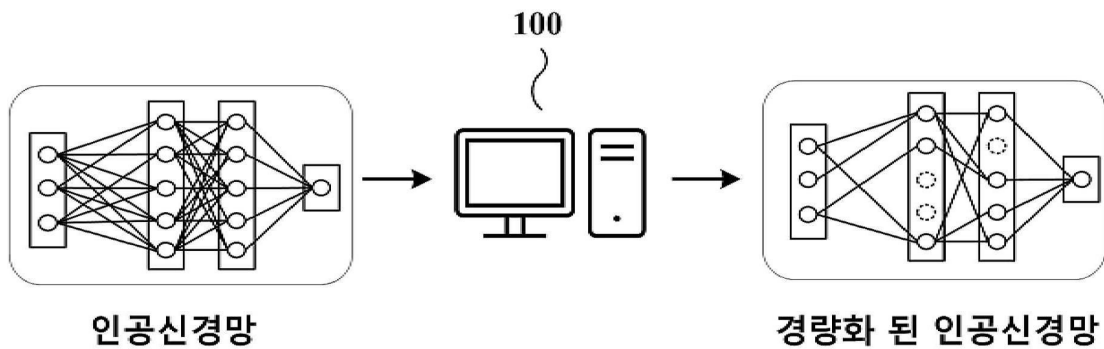
(54) 발명의 명칭 기여도를 이용한 인공신경망 경량화 방법 및 장치

(57) 요약

기여도를 이용한 인공신경망 경량화 방법은 경량화 장치가 인공신경망 모델 정보를 입력 받는 단계; 상기 경량화 장치가 상기 인공신경망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산하는 단계; 상기 경량화 장치가 상기 인공신경망의 각 계층(Layer) 마다 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 단계; 및 상기 경량화 장치가 상기 인공신경망의 각 계층마다 상기 선택된 링크를 제외한 나머지 링크의 가중치를 0으로 만드는 단계;를 포함한다.

상기 링크의 기여도는 새플리 값을 기반으로 계산된 값을 포함한다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호	1711152536
과제번호	2021-0-00739-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	방송통신산업기술개발
연구과제명	분산/협력 AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발
기 여 율	1/1
과제수행기관명	이화여자대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

경량화 장치가 인공지능망 모델 정보를 입력 받는 단계;

상기 경량화 장치가 상기 인공지능망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산하는 단계;

상기 경량화 장치가 상기 인공지능망의 각 계층(Layer)마다 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 단계; 및

상기 경량화 장치가 상기 인공지능망의 각 계층마다 상기 선택된 링크를 제외한 나머지 링크의 가중치를 0으로 만드는 단계;를 포함하되,

상기 인공지능망 모델은 순차적으로 나열된 복수개의 계층 및 상기 복수개의 계층 각각에 포함되며 링크를 통해 이전 계층과 다음 계층에 연결된 노드를 포함하고,

상기 링크의 기여도는 샐플리 값(Shapley Value)을 기반으로 계산된 값을 포함하는, 기여도를 이용한 인공지능망 경량화 방법.

청구항 2

제1항에 있어서,

상기 링크를 제거한 결과 다음 계층과 연결된 모든 링크가 제거된 노드가 있다면, 상기 경량화 장치가 상기 다음 계층과 연결된 모든 링크가 제거된 노드의 이전 계층에 연결된 모든 링크의 가중치를 0으로 만드는 단계;를 더 포함하는, 기여도를 이용한 인공지능망 경량화 방법.

청구항 3

제1항에 있어서,

상기 링크의 기여도는 아래의 수학적식을 이용해 계산된 값($\phi_{ij}^{(l)}$)을 포함하는, 기여도를 이용한 인공지능망 경량화 방법.

[수학적식]

$$\phi_{ij}^{(l)} = \sum_{J \subseteq I \setminus \{j\}} \frac{|J|!(|I| - |J| - 1)!}{I!} (v(J \cup \{j\}) - v(J))$$

상기 수학적식에서 $\phi_{ij}^{(l)}$ 는 l 번째 계층에서 i 번째 노드와 $(l-1)$ 번째 계층에서 j 번째 노드를 연결하는 링크의 기여도임. 상기 수학적식에서 I 는 l 번째 계층에서 링크들의 전체 집합을 의미함. 상기 수학적식에서 J 는 l 번째 계층 i 번째 노드와 $(l-1)$ 번째 계층의 j 번째 노드를 연결하는 링크를 제외한 나머지 링크들의 부분집합을 의미함. 상기 수학적식에서 $|I|$ 는 집합 I 의 원소의 개수를 의미함. 상기 수학적식에서 $|J|$ 는 집합 J 의 원소의 개수를 의미함. 상기 수학적식에서 $v(\cdot)$ 는 부분집합의 가치(value)를 의미함. 상기 부분집합의 가치는 Kernel SHAP방법에 기반하여 계산된 값을 포함함.

청구항 4

제1항에 있어서,

상기 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 단계는 상기 인공지능망 모델의 하나의 계층에 포함된 링크를 상기 링크의 기여도에 따라 내림 차순으로 나열하여 배열한 뒤, 아래 수학적식을 만

족하는 최소의 k^* 을 찾아 상기 배열의 앞에서 부터 순차적으로 k^* 개의 링크를 선택하는, 기여도를 이용한 인공신경망 경량화 방법.

[수학식]

$$\sum_{k=1}^{k^*} \Phi^{(l)}(k) \geq m \cdot \phi_T^{(l)}$$

상기 수학식에서 $\Phi^{(l)}(k)$ 은 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크에서 k 번째로 큰 링크의 기여도를 의미함. 상기 수학식에서 $\phi_T^{(l)}$ 는 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크의 기여도의 총합을 의미함. 상기 수학식에서 m 은 중요도 레벨을 의미함. 중요도 레벨 m 은 0과 1사이의 값을 가짐. 상기 수학식에서 k^* 는 자연수 값을 가짐.

청구항 5

인공신경망 모델 정보를 입력 받는 입력장치;

상기 인공신경망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산하고, 상기 인공신경망의 각 계층(Layer)마다 미리 설정된 기준을 만족하는 기여도를 가지는 링크를 적어도 하나 선택하고, 상기 인공신경망의 각 계층마다 상기 선택된 링크를 제외한 나머지 링크의 가중치를 0으로 만드는 연산장치; 및

상기 인공신경망 모델 정보를 저장하는 저장장치; 를 포함하되,

상기 인공신경망 모델은 순차적으로 나열된 복수개의 계층 및 상기 복수개의 계층 각각에 포함되며 링크를 통해 이전 계층과 다음 계층에 연결된 노드를 포함하고,

상기 링크의 기여도는 샐플리 값(Shapley Value)을 기반으로 계산된 값을 포함하는, 기여도를 이용한 인공신경망 경량화 장치.

청구항 6

제5항에 있어서,

상기 연산장치는 상기 링크를 제거한 결과 다음 계층(Layer)과 연결된 모든 링크가 제거된 노드가 있다면, 상기 다음 계층과 연결된 모든 링크가 제거된 노드의 이전 계층에 연결된 모든 링크의 가중치를 0으로 만드는, 기여도를 이용한 인공신경망 경량화 장치.

청구항 7

제5항에 있어서,

상기 링크의 기여도는 아래의 수학식을 이용해 계산된 값($\phi_{ij}^{(l)}$)을 포함하는, 기여도를 이용한 인공신경망 경량화 장치.

[수학식]

$$\phi_{ij}^{(l)} = \sum_{J \subseteq I \setminus \{j\}} \frac{|J|!(|I| - |J| - 1)!}{I!} (v(J \cup \{j\}) - v(J))$$

상기 수학식에서 $\phi_{ij}^{(l)}$ 는 l 번째 계층에서 i 번째 노드와 $(l-1)$ 번째 계층에서 j 번째 노드를 연결하는 링크의 기여도임. 상기 수학식에서 I 는 l 번째 계층에서 링크들의 전체 집합을 의미함. 상기 수학식에서 J 는 l 번째 계층 i 번째 노드와 $(l-1)$ 번째 계층의 j 번째 노드를 연결하는 링크를 제외한 나머지 링크들의 부분집합을 의미함. 상기

수학식에서 $|I|$ 는 집합 I 의 원소의 개수를 의미함. 상기 수학식에서 $|J|$ 는 집합 J 의 원소의 개수를 의미함. 상기 수학식에서 $v^{(i)}$ 는 부분집합의 가치(value)를 의미함. 상기 부분집합의 가치는 Kernel SHAP방법에 기반하여 계산된 값을 포함함.

청구항 8

제5항에 있어서,

상기 연산장치가 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 것은 상기 인공신경망 모델의 하나의 계층에 포함된 링크를 상기 링크의 기여도에 따라 내림 차순으로 나열하여 배열한 뒤, 아래 수학식을 만족하는 최소의 k^* 을 찾아 상기 배열의 앞에서 부터 순차적으로 k^* 개의 링크를 선택하는 것인, 기여도를 이용한 인공신경망 경량화 장치.

[수학식]

$$\sum_{k=1}^{k^*} \Phi^{(l)}(k) \geq m \cdot \phi_T^{(l)}$$

상기 수학식에서 $\Phi^{(l)}(k)$ 은 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크에서 k 번째로 큰 링크의 기여도를 의미함. 상기 수학식에서 $\phi_T^{(l)}$ 는 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크의 기여도의 총합을 의미함. 상기 수학식에서 m 은 중요도 레벨을 의미함. 중요도 레벨은 0과 1사이의 값을 가짐. 상기 수학식에서 k^* 는 자연수 값을 가짐.

청구항 9

제1항 내지 제4항 중 어느 하나의 항에 기재된 기여도를 이용한 인공신경망 경량화 방법을 실행하기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

발명의 설명

기술 분야

[0001] 이하 설명하는 기술은 기여도를 이용한 인공신경망 경량화 방법 및 장치에 대한 것이다.

배경 기술

[0002] 인공신경망은 다양한 분야에서 활용되고 있다.

[0003] 인공신경망이 실생활의 복잡한 문제를 해결하기 시작하면서, 인공신경망의 파라미터의 수도 증가하고 있다. 인공신경망의 파라미터 수가 많아지면, 모델의 크기와 복잡도는 증가한다. 이를 해결하기 위하여 다양한 모델을 경량화(lightweighting) 하는 기술들이 개발되고 있다.

선행기술문헌

특허문헌

[0004] (특허문헌 0001) 미국 공개특허공보 US 2022/0172050 A1

발명의 내용

해결하려는 과제

- [0005] 인공신경망을 경량화 하는 방법은 다양하다. 예를 들어 인공신경망을 경량화 하는 방법으로는 인공지능 모델을 압축하는 방법, 인공지능 모델을 최적화하는 방법, 인공지능 모델의 파라미터를 양자화하는 방법 등이 있다.
- [0006] 또는 인공신경망을 경량화 하는 방법에는 가지치기(프루닝, Pruning)이 포함된다. 가지치기 방법은 모델에서 중요하지 않은 파라미터나 레이어를 제거하여 모델의 크기를 줄이는 방법이다. 가지치기하여 인공신경망을 경량화 하는 방법에 있어서 중요하지 않은 파라미터나 레이어를 선택하는 과정은 중요하다.
- [0007] 이하 설명하는 기술은 샐플리-값(Shapley-value)을 기반으로 인공지능 모델에서 중요하지 않은 파라미터나 레이어를 파악하여 인공신경망을 경량화 할 수 있는 방법을 제공한다.

과제의 해결 수단

- [0008] 기여도를 이용한 인공신경망 경량화 방법은 경량화 장치가 인공신경망 모델 정보를 입력 받는 단계; 상기 경량화 장치가 상기 인공신경망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산하는 단계; 상기 경량화 장치가 상기 인공신경망의 각 계층(Layer) 마다 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 단계; 및 상기 경량화 장치가 상기 인공신경망의 각 계층마다 상기 선택된 링크를 제외한 나머지 링크의 가중치를 0으로 만드는 단계;를 포함한다.
- [0009] 상기 링크의 기여도는 샐플리 값을 기반으로 계산된 값을 포함한다.

발명의 효과

- [0010] 이하 설명하는 기술을 이용하면 중요도가 낮은 노드와 링크를 제거하여 인공신경망 모델을 경량화 시킬 수 있다.
- [0011] 이하 설명하는 기술을 이용하면 인공신경망 모델을 경량화 시키면서도 성능을 향상시키거나 유지할 수 있다.
- [0012] 이하 설명하는 기술을 이용하면 목표 희소레벨(Sparsity level)에 맞는 중요도 레벨(Importance level)를 설정하여 인공신경망 모델을 경량화 할 수 있다.

도면의 간단한 설명

- [0013] 도1은 경량화 장치가 기여도를 이용하여 인공신경망을 경량화하는 전체적인 과정의 예이다.
- 도2는 기여도를 이용하여 인공신경망을 경량화 하는 구체적인 과정의 예이다.
- 도3은 기여도를 이용한 인공신경망 경량화 방법을 컴퓨터상에서 실행하는데 필요한 알고리즘을 요약한 것이다.
- 도4는 링크의 기여도의 예시를 보여준다.
- 도5는 합성 데이터 셋을 이용해 실험할 때 이용된 오토인코더 구조이다.
- 도6은 실험에 이용된 합성 데이터 셋의 특징을 보여준다.
- 도7은 합성 데이터를 이용해 오토인코더를 학습한 후, 인코더와 디코더의 가중치 행렬을 보여준다.
- 도8은 중요도 레벨을 0.8로 설정하여 오토인코더를 경량화 하였을 때, 인코더와 디코더의 가중치 행렬을 보여준다.
- 도9는 경량화 과정에서 계산된 링크의 중요도를 보여준다.
- 도10은 경량화 과정에서 계산된 노드의 중요도 및 링크의 중요도의 분산(Variance)을 보여준다.
- 도11은 중요도 레벨에 따른 경량화 된 오토인코더의 출력 값이다.
- 도12및 도13은 경량화 정도에 따른 오토인코더의 출력 값을 분류모델에 입력하였을 때, 분류모델의 성능을 평가한 결과이다.
- 도14 및 도15는 경량화 된 오토인코더 모델과 종래 기술을 비교한 결과이다.
- 도16은 경량화 장치의 구성의 예이다.

발명을 실시하기 위한 구체적인 내용

- [0014] 이하 설명하는 기술은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있다. 명세서의 도면에 이하 설명하는 기술의 특정 실시 형태가 기재될 수 있다. 그러나, 이는 이하 설명하는 기술의 설명을 위한 것이며 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니다. 따라서 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경 물, 균등 물 내지 대체 물이 이하 설명하는 기술에 포함하는 것으로 이해되어야 한다.
- [0015] 이하 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 기재된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계 동작 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.
- [0016] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성 부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성 부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.
- [0017] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.
- [0018] 이하 설명하는 기술에서 인공신경망 모델의 경량화는 가지치기에 해당한다. 경량화는 인공신경망을 구성하는 노드와 링크 중 일부를 제거하는 것을 포함할 수 있다. 노드와 링크를 제거하기 위하여 해당 링크의 가중치를 0으로 만들 수 있다. 가중치를 0으로 만드는 것을 비활성화(Deactivated)라고 부를 수도 있다.
- [0019] 이하 설명하는 기술에서 인공신경망 모델은 인간 뇌 신경망을 모방하여 만든 모델이다. 이하 인공신경망은 적어도 하나의 입력 계층, 적어도 하나의 은닉 계층 및 적어도 하나의 출력 계층을 포함할 수 있다. 출력 계층은 분류기 등일 수 있다.
- [0020] 노드는 유닛(Unit)이라고 부를 수도 있다. 링크는 엣지(Edge)라고 부를 수도 있다. 인공신경망 모델은 복수개의 파라미터를 포함할 수 있다. 파라미터는 가중치(Weight)와 편향(Bias)을 포함할 수 있다.
- [0021] 연구자는 오토인코더(autoencoder)에 경량화 기법을 적용하였다. 따라서, 이하 설명 중 일부는 오토인코더를 대상으로 한다. 다만 이하 설명하는 기술은 오토인코더와 같은 특정 구조의 모델에 제한되지 않고, MLP(multi-layer perceptron) 등과 같이 다수의 계층을 갖는 다양한 인공신경망 모델에 적용될 수 있다.
- [0023] 먼저 인공신경망 경량화 장치(이하, 경량화 장치)가 기여도(Contribution)를 이용하여 인공신경망을 경량화 하는 전체적인 과정을 설명한다.
- [0024] 도1은 경량화 장치(100)가 기여도를 이용한 인공신경망을 경량화 하는 전체적인 과정이다.
- [0025] 경량화 장치(100)는 PC, 노트북, 스마트기기, 서버 또는 데이터처리 전용 칩셋 등의 형태를 가질 수 있다
- [0026] 경량화 장치(100)는 인공신경망 모델 정보를 입력 받을 수 있다. 경량화 장치(100)는 인공신경망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산할 수 있다. 경량화 장치(100)는 인공신경망의 각 계층(Layer) 마다 미리 설정된 기준을 만족하는 기여도를 가지는 링크를 적어도 하나 선택할 수 있다. 경량화 장치(100)는 인공신경망의 각 계층마다 선택된 링크를 제외한 나머지 링크를 제거할 수 있다.
- [0027] 링크를 제거한 결과 다음 계층과 연결된 모든 링크가 제거된 노드가 있다면, 경량화 장치(100)는 상기 노드를 제거할 수 있다. 구체적으로 경량화 장치(100)는 다음 계층과 연결된 모든 링크가 제거된 노드가 있는 경우, 해당 노드와 이전 계층을 연결하는 모든 링크를 제거할 수 있다.
- [0028] 링크의 기여도는 각 링크가 해당 계층의 출력에 주는 영향을 의미할 수 있다. 링크의 기여도는 샐플리 값(Shapley Value)을 기반으로 계산된 값일 수 있다.

- [0030] 이하 경량화 장치가 기여도를 이용해서 인공신경망을 경량화 하는 과정을 구체적으로 설명한다.
- [0031] 도 2는 경량화 장치가 기여도를 이용한 인공신경망 경량화 방법을 수행하는 순서도(200)이다. 도3은 기여도를 이용한 인공신경망 경량화 방법을 컴퓨터상에서 실행하는데 필요한 알고리즘을 나타낸다.
- [0032] 경량화 장치는 인공신경망 모델 정보를 입력 받을 수 있다(210)
- [0033] 인공신경망 모델은 순차적으로 나열된 복수개의 계층 및 상기 복수개의 계층 각각에 포함되며 링크를 통해 이전 계층과 다음 계층에 연결된 노드를 포함할 수 있다.
- [0034] 인공신경망 모델 정보는 인공신경망의 구조 및 인공신경망의 파라미터 값 등의 정보를 포함한다. 예를 들어 인공신경망 모델 정보는 인공신경망 계층의 개수, 각 계층에 포함된 노드들의 개수, 각 노드를 연결하는 링크의 형태 등의 정보를 포함할 수 있다. 또는 인공신경망 모델 정보는 노드 및 링크의 파라미터 값(가중치 및 편향 등)을 포함할 수 있다.
- [0035] 경량화 장치는 인공신경망 모델의 각 계층별로 모든 링크의 기여도를 계산할 수 있다(220).
- [0036] 링크의 기여도는 각 링크가 해당 계층의 출력에 주는 영향을 의미할 수 있다. 링크의 기여도는 링크의 중요도 (Link Importance, LI)라고 부를 수 있다.
- [0037] 링크의 기여도는 새폴리 값을 기반으로 계산된 값일 수 있다.
- [0038] 수학식 1은 링크의 기여도를 계산하는데 이용되는 식이다.
- [0039] 구체적으로 수학식1은 l 번째 계층의 i 번째 노드와 $(l-1)$ 번째 계층의 j 노드를 연결하는 링크의 기여도 ($\phi_{ij}^{(l)}$)를 계산하는데 이용되는 식이다.

수학식 1

$$\phi_{ij}^{(l)} = \sum_{J \subseteq I \setminus \{j\}} \frac{|J|!(|I| - |J| - 1)!}{I!} (v(J \cup \{j\}) - v(J))$$

- [0040]
- [0041] 수학식1에서 I 는 l 번째 계층에서 링크들의 전체 집합을 의미한다. 수학식1에서 J 는 l 번째 계층 i 번째 노드와 $(l-1)$ 번째 계층의 j 번째 노드를 연결하는 링크를 제외한 나머지 링크들의 부분집합을 의미한다. 수학식 1에서 $|I|$ 는 집합 I 의 원소의 개수를 의미한다. 수학식1에서 $|J|$ 는 집합 J 의 원소의 개수를 의미한다. 수학식1에서 $v(\cdot)$ 는 부분집합의 가치(value)를 의미한다. 부분집합의 가치 값은 Kernel SHAP방법에 기반하여 계산된 값이다. Kernel SHAP는 논문 (A Unified Approach to Interpreting Model Predictions, Lundberg & Lee, 2017) 등에서 소개된 것일 수 있다.
- [0042] 도 4는 $l=1$ 일 때 수학식 1로 계산되는 링크의 기여도의 예시를 보여준다.
- [0043] 경량화 장치는 인공신경망의 각 계층 마다 미리 설정된 기준을 만족하는 기여도를 가지는 링크를 적어도 하나 선택할 수 있다(230).
- [0044] 미리 설정된 기준을 만족하는 링크를 적어도 하나 선택하는 것은 미리 설정된 임계값 이상의 기여도를 가지는 링크를 적어도 하나 선택하는 것일 수 있다.
- [0045] 상기 미리 설정된 기준을 만족하는 기여도를 가지는 링크를 적어도 하나 선택하는 것은 상기 인공신경망 모델의 하나의 계층에 포함된 링크를 상기 링크의 기여도에 따라 내림 차순으로 나열하여 배열한 뒤, 수학식3을 만족하는 최소의 k^* 을 찾아 상기 배열의 앞부터 순차적으로 k^* 개의 링크를 선택하는 것일 수 있다.
- [0046] 수학식2는 l 번째 계층의 링크의 기여도를 내림차순($\Phi^{(l)}(k) \geq \Phi^{(l)}(k+1)$)으로 나열한 배열 ($\Phi^{(l)}$)를 보여준다.

수학식 2

$$\Phi^{(l)} = [\Phi^{(l)}(1), \Phi^{(l)}(2), \dots, \Phi^{(l)}(n^{(l-1)}n^{(l)})]$$

[0047]

[0048] 수학식 2에서 $n^{(l)}$ 은 l 번째 계층에 포함된 노드의 총 개수이다. 수학식2의 배열($\Phi^{(l)}$)은 총 ($n^{(l)}n^{(l-1)}$) 개의 값이 포함되어 있다. 수학식2에서 $\Phi^{(l)}(k)$ 는 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크 중에서 k 번째로 기여도가 큰 링크의 기여도를 의미한다.

[0049] 수학식3은 중요도 레벨(m)을 기초로 기여도가 높은 링크를 선택할 때 이용하는 식이다. 구체적으로 수학식3은 중요도 레벨(m)을 기초로 l 번째 계층에서 기여도가 높은 링크를 선택한 뒤 생성한 집합 ($\Gamma^{(l)}$)에 대한 것이다.

수학식 3

$$\Gamma^{(l)} = \left\{ (i, j) \left| \sum_{k=1}^{k^*} \Phi^{(l)}(k) \geq m \cdot \phi_T^{(l)} \right. \right\}$$

[0050]

[0051] 수학식3에서 $\phi_T^{(l)}$ 는 l 번째 계층과 $(l-1)$ 번째 계층을 연결하는 링크의 중요도의 총합을 의미한다. $\phi_T^{(l)}$ 는 수학식 4을 이용해 계산된다.

수학식 4

$$\phi_T^{(l)} = \sum_{i=1}^{n^{(l)}} \sum_{j=1}^{n^{(l-1)}} \phi_{ij}^{(l)}$$

[0052]

[0053] 수학식 3에서 m 은 중요도 레벨을 의미한다. 중요도 레벨 m 은 0과 1사이의 값을 가진다.

[0054] 수학식 3에서 k^* 는 조건을 만족하는 최소의 자연수일 수 있다.

[0055] 수학식 3의 집합 $\Gamma^{(l)}$ 는 k^* 개의 순서쌍을 가진다. 각 순서쌍은 l 번째 계층의 i 번째 노드와 $(l-1)$ 번째 계층의 j 번째 노드를 연결하는 링크에 대한 정보를 가진다. 예를 들어 (2, 3)은 l 번째 계층의 2번째 노드와 $(l-1)$ 번째 계층의 3번째 노드를 연결하는 링크에 대응하는 순서쌍에 대한 것이다.

[0056] 중요도 레벨이 작아질수록 집합($\Gamma^{(l)}$)에 속하는 순서쌍의 수가 적어진다. 후술하듯이 집합($\Gamma^{(l)}$)에 속하는 순서쌍의 수가 적어질수록 제거되는 링크의 수가 증가한다.

[0057] 경량화 장치는 인공신경망의 각 계층 마다 선택된 링크를 제외한 나머지 링크를 제거할 수 있다(240). 링크를 제거하는 것은 링크의 가중치를 0으로 하는 것을 포함할 수 있다. 이 과정을 희소화(Sparsification) 과정이라고 할 수 있다.

[0058] 경량화 장치는 마스킹 함수(Masking Function, $\mathcal{M}$)를 이용해 희소화 과정을 수행할 수 있다.

[0059] 마스킹 함수는 선택된 링크를 제외한 나머지 링크의 가중치를 0으로 만들 수 있다. 마스킹 함수는 인공신경망 모델의 가중치 행렬($\mathbf{W}, \mathbf{W}'$)을 입력 받아 기여도가 작은 링크가 제거된 희소 가중치 행렬($\mathbf{W}^*, \mathbf{W}'^*$)을 출력한다.

[0060] 수학식5은 마스크 함수를 통해 최소 가중치 행렬을 생성하는데 이용되는 식이다.

수학식 5

$$\mathbf{W}^* = \mathcal{M}(\mathbf{W})$$

$$\mathbf{W}'^* = \mathcal{M}(\mathbf{W}')$$

[0061]

[0062] 구체적으로 마스크 함수는 집합 $\Gamma^{(l)}$ 에 포함되지 않는 순서쌍에 대응되는 링크의 가중치를 0으로 만든다. 반대로 마스크 함수는 $\Gamma^{(l)}$ 에 포함된 순서쌍에 대응되는 링크의 가중치는 그대로 둔다.

[0063] 수학식6은 마스크 함수가 집합 $\Gamma^{(l)}$ 에 포함되지 않는 순서쌍에 상응하는 링크의 가중치를 0으로 만드는데 이용하는 식이다.

수학식 6

$$w_{ij}^{*(l)} = 0 \text{ if } \phi_{ij}^{(l)} \in \Gamma^{(l)c}$$

[0064]

[0065] 더 나아가 경량화 장치는 다음 계층과 연결된 모든 링크가 제거된 노드가 있는 경우, 해당 노드와 이전 계층을 연결하는 모든 링크도 제거할 수 있다(250). 다음 계층과 연결된 모든 링크가 제거된 노드는 불필요함에도, 이전 계층에서 불필요한 연산이 수행될 수 있기 때문이다.

[0066] 즉 수학식7의 조건을 만족하면 l 번째 계층의 k 번째 노드에 연결된 링크는 제거된다.

수학식 7

$$(1) \quad w_{kj}^{(l)} \neq 0, \quad 1 \leq j \leq n^{(l-1)}$$

$$(2) \quad w_{ik}^{(l+1)} = 0, \quad 1 \leq i \leq n^{(l)}$$

[0067]

[0068] 또한 경량화 장치는 희소레벨(Sparsity Level, η)을 계산할 수 있다(미도시).

[0069] 희소레벨은 인공신경망 모델이 얼마나 경량화 되었는지를 보여주는 수치일 수 있다. 구체적으로 희소레벨은 인공신경망 모델에서 링크가 얼마나 제거되었는지를 보여주는 수치일 수 있다.

[0070] 희소레벨이 높을수록 인공신경망 모델은 경량화가 많이 되었다는 것을 알 수 있다. 반대로 희소레벨이 낮을수록 인공신경망 모델은 경량화가 적게되었다는 것을 알 수 있다.

[0071] 수학식 8은 희소레벨(η)을 계산하는데 이용되는 식이다.

수학식 8

$$\eta = \frac{\sum_{l=1}^{2L-1} |\Gamma^{(l)c}|}{\sum_{l=1}^{2L-1} |\Gamma^{(l)} \cup \Gamma^{(l)c}|}$$

[0072]

- [0073] 또한 경량화 장치는 인공신경망 모델의 각 노드의 기여도를 계산할 수 있다(미도시).
- [0074] 노드의 기여도는 각 노드가 다음 계층의 출력에 얼마나 영향을 주는지를 의미할 수 있다. 노드의 기여도는 노드의 중요도(Unit Importance, UI)라고 부를 수도 있다.
- [0075] 수학적식 9는 노드의 기여도를 계산하는데 이용되는 식이다.
- [0076] 구체적으로 수학적식9는 $(l-1)$ 번째 계층에서 l 번째 노드의 중요도를 계산하는데 이용되는 식이다. 노드의 중요도는 수학적식9와 같이 $(l-1)$ 번째 계층에서 l 번째 노드와 연결된 l 번째 계층의 모든 노드와의 링크의 기여도의 평균으로 정의된다.

수학적식 9

$$\bar{v}_j^{(l-1)} = \frac{1}{n^{(l)}} \sum_{i=1}^{n^{(l)}} |\phi_{ij}^{(l)}|$$

- [0077]
- [0078] 수학적식9에서 $n^{(l)}$ 는 l 번째 계층의 노드의 총 개수를 의미한다.
- [0080] 연구자는 전술한 기여도를 이용한 인공신경망 경량화 방법을 이용해 실제 인공신경망을 경량화 하였다. 연구자가 경량화 한 인공신경망 모델은 오토인코더(AutoEncoder)이다. 연구자는 경량화 된 인공신경망의 성능을 평가하는 실험을 진행하였다. 연구자는 성능 평가를 위해 합성 데이터셋(Synthetic Dataset) 및 실제 데이터(Real Dataset)를 이용하였다.
- [0081] 이하 연구자가 실제 인공신경망을 경량화 하고, 성능을 평가한 결과를 살펴본다.
- [0083] 먼저 도 5 내지 도 10을 통해 합성 데이터를 이용해 실험한 결과를 설명한다.
- [0085] 도 5는 합성 데이터를 이용해 실험할 때 이용된 오토인코더 구조이다.
- [0086] 실험에 이용된 오토인코더는 입력계층, 출력계층 및 은닉계층을 포함한다.
- [0087] 입력계층은 6개의 노드를 포함한다. 은닉계층은 3개의 노드를 포함한다. 출력계층은 6개의 노드를 포함한다.
- [0088] 입력 데이터 $\mathbf{x}$ 는 6차원으로 이루어져 있다. 각 차원의 데이터 집합(x_1 내지 x_6)은 입력계층의 각 노드에 입력된다. 은닉층의 데이터 $\mathbf{z}$ 는 3차원으로 이루어져있으며 z_1 내지 z_3 를 포함한다. 출력 데이터 $\hat{\mathbf{x}}$ 은 6차원으로 이루어져 있으며, 각 차원의 데이터 집합 $\hat{x}_1$ 내지 $\hat{x}_6$ 을 포함한다.
- [0089] 인코더(Encoder)의 가중치 행렬($\mathbf{W}$)은 6 X 3 행렬로 구성된다. 디코더(Decoder)의 가중치 행렬($\mathbf{W}'$)은 3 X 6 행렬로 구성된다. 활성화 함수(Activation Function)는 ReLu이다.
- [0091] 도 6은 실험에 이용된 합성 데이터셋의 특징을 보여준다. 구체적으로 도 6은 오토인코더 입력계층의 6개 노드에 입력되는 각 차원의 데이터 집합에 속하는 값의 히스토그램이다.
- [0092] 오토인코더 입력계층의 첫번째 노드에는 일정한 상수 값(constant)의 1/2이 입력된다. 즉 x_1 은 일정한 상수 값으로만 이루어져 있다.
- [0093] 오토인코더 입력계층의 두번째 노드에는 0부터 1까지 균일한 분포(uniform distribution)의 값이 입력된다.

- [0094] 오토인코더 입력계층의 세번째 노드에는 2개의 클래스로부터 샘플링 된 데이터가 입력된다. 2개의 클래스 중 하나는 0 부터 1/4 사이 값이 균일하게 분포되어 있다. 2개의 클래스 중 나머지 하나는 3/4 부터 1 사이 값이 균일하게 분포되어 있다.
- [0095] 오토인코더 입력계층의 네번째 노드에는 2개의 클래스로부터 샘플링 된 데이터가 입력된다. 2개의 클래스 중 하나는 평균이 1/4이고 분산이 0.1인 가우스 분포에 따라 값이 분포되어 있다. 2개의 클래스 중 나머지 하나는 평균이 3/4이고 분산이 0.1인 가우스 분포에 따라 값이 분포되어 있다.
- [0096] 오토인코더 입력계층의 다섯번째 노드에는 첫 번째 노드와 두 번째 노드에 들어가는 입력 값의 합(x_1+x_2)이 입력된다.
- [0097] 오토인코더 입력계층의 여섯번째 노드에는 세 번째 노드와 네 번째 노드에 들어가는 입력 값의 합(x_3+x_4)이 입력된다.
- [0099] 도 7은 합성 데이터를 이용해 오토인코더를 학습한 후, 인코더와 디코더의 가중치 행렬($\mathbf{W}$, $\mathbf{W}'$)을 보여준다.
- [0100] 도 7에서 확인할 수 있듯이 최소화 이전의 가중치 행렬의 모든 값은 0이 아니다. 따라서 단순히 학습단계가 완료된 경우 모든 링크에 대하여 계산하여야 하기 때문에 계산 량이 많다.
- [0102] 연구자는 중요도 레벨(m)을 0.8로 설정하여 도7의 오토인코더를 경량화 하였다.
- [0103] 입력계층과 은닉계층을 연결하는 링크 중 기여도가 높은 링크를 선택하여 생성한 집합($\Gamma^{(1)}$)에 속하는 순서쌍은 (1, 2), (1, 4), (1, 5), (1, 6), (2, 3), (2, 5), (2, 6) 및 (3, 5)이다. 이때 은닉층의 두 번째 노드와 출력계층의 노드를 연결하는 모든 링크가 제거된 것을 확인할 수 있다. 이에 은닉층의 두 번째 노드는 제거되었다(도2의 250과정). 따라서 순서쌍 (2,3), (2,5) 및 (2,6)은 $\Gamma^{(1)}$ 에 포함시킬 이유가 없다. 따라서 최종적으로 $\Gamma^{(1)}$ 에 포함된 순서쌍은 (1, 2), (1, 4), (1, 5), (1, 6) 및 (3, 5)이다. 은닉계층과 출력계층을 연결하는 링크 중 기여도가 높은 링크를 선택하여 생성한 집합 $\Gamma^{(2)}$ 에 속하는 순서쌍은 (2, 1), (2, 3) (3, 1), (3, 2), (4, 1), (5, 3) 및 (6, 1) 이다.
- [0104] 중요도 레벨을 0.8로 하였을 때 계산된 최소레벨은 0.67이다(수학식 10).
- 수학식 10**
- $$\eta = \frac{|\Gamma^{(1)c}| + |\Gamma^{(2)c}|}{|\Gamma^{(1)} \cup \Gamma^{(1)c}| + |\Gamma^{(2)} \cup \Gamma^{(2)c}|} = \frac{13 + 11}{18 + 18} = 0.67$$
- [0105]
- [0106] 도 8은 중요도 레벨을 0.8로 설정하여 오토인코더를 경량화 하였을 때의 가중치 행렬($\mathbf{W}^*$, $\mathbf{W}'^*$)을 보여준다. 도 8에서 점선으로 표시된 링크는 경량화 과정에서 제거된 링크이다.
- [0107] 도 8에서 확인할 수 있듯이 가중치 행렬의 일부 값이 0이다. 가중치가 0인 링크는 연산과정에서 계산되지 아니한다. 따라서 이를 통해 계산량을 줄일 수 있다.
- [0109] 도9는 경량화 과정에서 계산된 링크의 중요도(Link Importance, LI)을 보여준다.
- [0110] 도10은 경량화 과정에서 계산된 노드의 중요도(Unit Importance, UI) 및 링크의 중요도(LI)의 분산(Variance)을 보여준다.
- [0111] 도10의(a)는 수학식9를 통해 계산된 노드의 중요도 값이다.

- [0112] 도10의(b)는 입력 데이터의 분포가 노드와 링크의 중요도에 주는 영향을 분석하기 위하여 계산된 링크의 중요도 분산(Variance of Link Importance)이다. 링크의 중요도 분산이 높을수록 해당 노드가 다음 계층 노드 중 일부에 집중적으로 영향을 준다. 링크의 중요도 분산이 낮을수록 해당 노드가 다음 계층 노드 전체에 균일하게 영향을 준다.
- [0113] x5는 x1와 x2의 합이고, x6는 x3및 x4의 합이다. 따라서 x5와 x6은 x1내지x4 의 정보를 포함할 수 있다. 그리하여 학습과정에서 x5와 x6은 결과 값에 더 많이 기여할 수 있다. 반대로 학습과정에서 x1 내지 x4는 결과 값에 더 적은 기여를 할 수 있다. 따라서 도10에서 x5와 x6가 입력되는 노드의 중요도가 다른 노드(x1 내지 x4가 입력되는 노드)의 중요도보다 높은 것을 확인할 수 있다.
- [0114] x1는 상수 값으로만 이루어져있다. 따라서 학습과정에서 x1은 결과값에 기여를 하지 못한다. 이에 도10에서 x1이 입력되는 첫 번째 노드의 중요도는 0인 것을 확인할 수 있다. 즉 x1은 더미(dummy) 데이터로 취급될 수 있다.
- [0115] x5와 x6가 입력되는 노드는 유사한 노드의 중요도(UI)을 가진다. 그러나 x5가 입력되는 노드에 연결된 링크의 중요도 분산(0.0013)은 x6가 입력되는 노드에 연결된 링크의 중요도 분산(0.0095)보다 상대적으로 작다. 이는 도6에서 확인할 수 있듯이 x5는 전체적으로 일정한 분포를 따르기 때문이다. 따라서 즉 x5는 은닉 계층의 각 노드에 균일하게 영향을 준다. 이는 도9에서 확인할 수 있다.
- [0116] 반대로 x3,x4 및 x6가 입력되는 노드에 연결된 링크의 중요도 분산은 상대적으로 높다(0.0027, 0.0012, 0.0095). 이는 x3,x4 및 x6은 은닉 계층의 노드 중 일부에 집중된 영향을 준다. 이는 도9에서 확인할 수 있다.
- [0118] 이하 도 11 내지 도 14을 통해 실제 데이터를 이용해 실험한 결과를 살펴본다. 실험에 이용된 실제 데이터는 MNIST 데이터 셋이다.
- [0120] 실제 데이터를 이용해 실험하기 위해서 이용된 오토인코더는 도 5 내지 도 10에서 이용된 오토인코더와 다르다. 도 11 내지 도 14에서의 오토인코더는 입력계층, 은닉계층, 출력계층을 포함한다. 입력계층은 784개의 노드를 포함한다. 은닉계층은 32개의 노드를 포함한다. 출력계층은 784개의 노드를 포함한다. 이용된 활성화함수는 LeakyReLU이다. 마지막 계층에서 이용된 활성화함수는 Sigmoid이다.
- [0122] 도11은 중요도 레벨에 따른 경량화 된 오토인코더의 출력 값이다.
- [0123] 도 11에서 회색으로 표시된 부분은 경량화 된 오토인코더의 연산에서 제외되어, 복원되지 않은 부분이다. 중요도 레벨(m)이 낮을수록 경량화 된 오토인코더가 입력 이미지의 가장자리 부분을 잘 복원하지 못하는 것을 확인할 수 있다. 중요도 레벨이 낮을수록 오토인코더의 기여도가 낮은 링크가 더 많이 제거되기 때문이다.
- [0125] 도 12는 및 도 13은 경량화 정도에 따른 오토인코더의 출력 값을 분류모델에 입력하였을 때, 분류모델의 성능을 평가한 결과이다. 도 12 및 도 13에서 이용된 분류모델은 Gaussian radial basis 함수에 기초한 SVM(Support Vector Machine)이다.
- [0126] 도 12 및 도 13을 통해 희소화 레벨이 약 0.8이 될 정도로 경량화가 진행되는 과정에서, 분류성능이 증가하는 것을 확인할 수 있다. 이는 초기 경량화 과정에서 Occam's hill에 따른 학습된 노이즈가 제거되기 때문일 수 있다.
- [0127] 도 12 및 도 13을 통해 희소화 레벨이 0.9가 될 정도로 경량화가 진행된 경우에도 희소화 이전에 비해 성능이 오직 1.37%만 감소된 것을 확인할 수 있다.
- [0128] 도 12 및 도 13을 통해 경량화가 진행되면 성능이 향상되거나, 급격히 하락하지 않는 것을 확인할 수 있다.
- [0130] 도 14 및 도 15는 경량화 된 오토인코더 모델과 종래 기술과 비교한 결과이다.

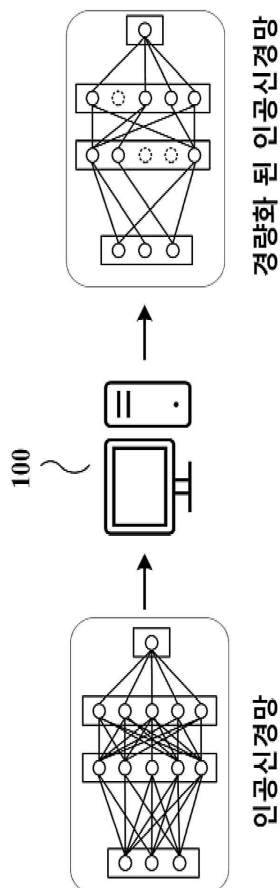
- [0131] 구체적으로 도 14 및 도 15는 오토인코더의 출력 값을 분류모델에 입력하였을 때 분류모델의 성능을 평가한 결과이다. 비교를 위해 이용된 종래 기술은 기술인 드랍아웃(Dropout) 또는 k-희소(k-sparse)에 기반한 경량화 방법이 적용된 오토인코더이다.
- [0132] 도 14 및 도 15를 통해 경량화 된 오토인코더 모델을 이용하면 종래 기술과 비교할 때 분류성능이 높은 것을 확인할 수 있다. 또한 경량화 된 오토인코더 모델을 이용하는 경우에 경량화 정도가 커지더라도 분류성능이 급격하게 하락하지 않는 것을 확인할 수 있다. 특히 희소화 레벨이 0.69가 되는 경우 경량화 된 오토인코더 모델의 성능이 가장 좋은 것을 확인할 수 있다.
- [0134] 이하 경량화 장치의 구성에 대하여 설명한다.
- [0135] 도16은 경량화 장치의 구성의 예이다.
- [0136] 경량화 장치(300)는 도1에서 설명한 경량화 장치(100)에 해당할 수 있다.
- [0137] 경량화 장치(300)는 물리적으로 다양한 형태로 구현될 수 있다. 예를 들어 경량화 장치(300)는 PC, 노트북, 스마트폰, 서버 또는 데이터처리 전용 칩셋 등의 형태를 가질 수 있다.
- [0138] 경량화 장치(300)는 입력장치(310), 저장장치(320), 연산장치(330), 출력장치(340), 인터페이스 장치(350) 및 통신장치(360)를 포함할 수 있다.
- [0139] 입력장치(310)는 일정한 명령 또는 데이터를 입력 받는 인터페이스 장치(키보드, 마우스, 터치스크린 등)를 포함할 수도 있다. 입력장치(310)는 별도의 저장장치(USB, CD, 하드디스크 등)를 통하여 정보를 입력 받는 구성을 포함할 수도 있다. 입력장치(310)는 입력 받는 데이터를 별도의 측정장치를 통하여 입력 받거나, 별도의 DB를 통하여 입력 받을 수도 있다. 입력장치(310)는 유선 또는 무선 통신을 통해 데이터를 입력 받을 수 있다.
- [0140] 입력장치(310)는 기여도를 이용한 인공신경망 경량화 방법을 수행하는데 필요한 정보 및 모델을 입력 받을 수 있다. 입력장치(310)는 인공신경망 모델 정보를 입력 받을 수 있다.
- [0141] 저장장치(320)는 입력장치(310)을 통해 입력 받은 정보를 저장할 수 있다. 저장장치(320)는 연산장치(330)가 연산하는 과정에서 생성되는 정보를 저장할 수 있다. 즉 저장장치(320)는 메모리를 포함할 수 있다. 저장장치(320)는 연산장치(330)가 계산한 결과를 저장할 수 있다.
- [0142] 저장장치(320)는 기여도를 이용한 인공신경망 경량화 방법을 수행하는데 필요한 정보 및 모델을 저장할 수 있다. 저장장치(320)는 인공신경망 모델 정보를 저장할 수 있다.
- [0143] 연산장치(330)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드 된 칩과 같은 장치일 수 있다. 연산장치(330)는 경량화 장치를 제어하는 제어신호를 생성할 수 있다.
- [0144] 연산장치(330)는 기여도를 이용한 인공신경망 경량화 방법을 수행하는데 필요한 연산을 수행할 수 있다. 연산장치(330)는 인공신경망 모델의 각 링크의 기여도를 계산할 수 있다. 연산장치(330)는 상기 인공신경망 모델의 각 계층별로 모든 링크(Link)의 기여도를 계산할 수 있다. 연산장치(330)는 상기 인공신경망의 각 계층(Layer)마다 미리 설정된 기준을 만족하는 기여도를 가지는 링크를 적어도 하나 선택할 수 있다. 연산장치(330)는 인공신경망의 각 계층마다 상기 선택된 링크를 제외한 나머지 링크를 제거할 수 있다. 링크를 제거한 결과 다음 계층과 연결된 모든 링크가 제거된 노드가 있다면, 연산장치(330)는 다음 계층과 연결된 모든 링크가 제거된 노드의 이전 계층에 연결된 모든 링크의 가중치를 0으로 만들 수 있다. 연산장치(330)는 희소레벨을 계산할 수 있다. 연산장치(330)는 인공신경망 모델의 각 노드의 기여도를 계산할 수 있다.
- [0145] 출력장치(340)는 일정한 정보를 출력하는 장치가 될 수도 있다. 출력장치(340)는 데이터 과정에 필요한 인터페이스, 입력된 데이터, 분석결과 등을 출력할 수도 있다. 출력장치(340)는 디스플레이, 문서를 출력하는 장치, 등과 같이 물리적으로 다양한 형태로 구현될 수도 있다.
- [0146] 인터페이스 장치(350)는 외부로부터 일정한 명령 및 데이터를 입력 받는 장치일 수 있다. 인터페이스 장치(350)는 물리적으로 연결된 입력 장치 또는 외부 저장장치로부터 인공신경망 모델 정보를 입력 받을 수 있다. 인터페이스 장치(350)는 경량화 장치(300)를 제어하기 위한 제어신호를 입력 받을 수 있다. 인터페이스 장치(350)는 경량화 장치(300)가 경량화 한 결과를 출력할 수 있다.
- [0147] 통신장치(360)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미할 수 있다. 통

신장치(360)는 경량화 장치(300)를 제어하는데 필요한 제어 신호를 수신할 수 있다. 통신장치(360)는 경량화 장치(300)가 분석한 결과를 전송할 수 있다.

- [0149] 전술한 기여도를 이용한 인공신경망 경량화 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다.
- [0150] 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.
- [0151] 비일시적 판독 가능 매체는 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM (read-only memory), PROM (programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.
- [0152] 일시적 판독 가능 매체는 스테틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(Synclink DRAM, SDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

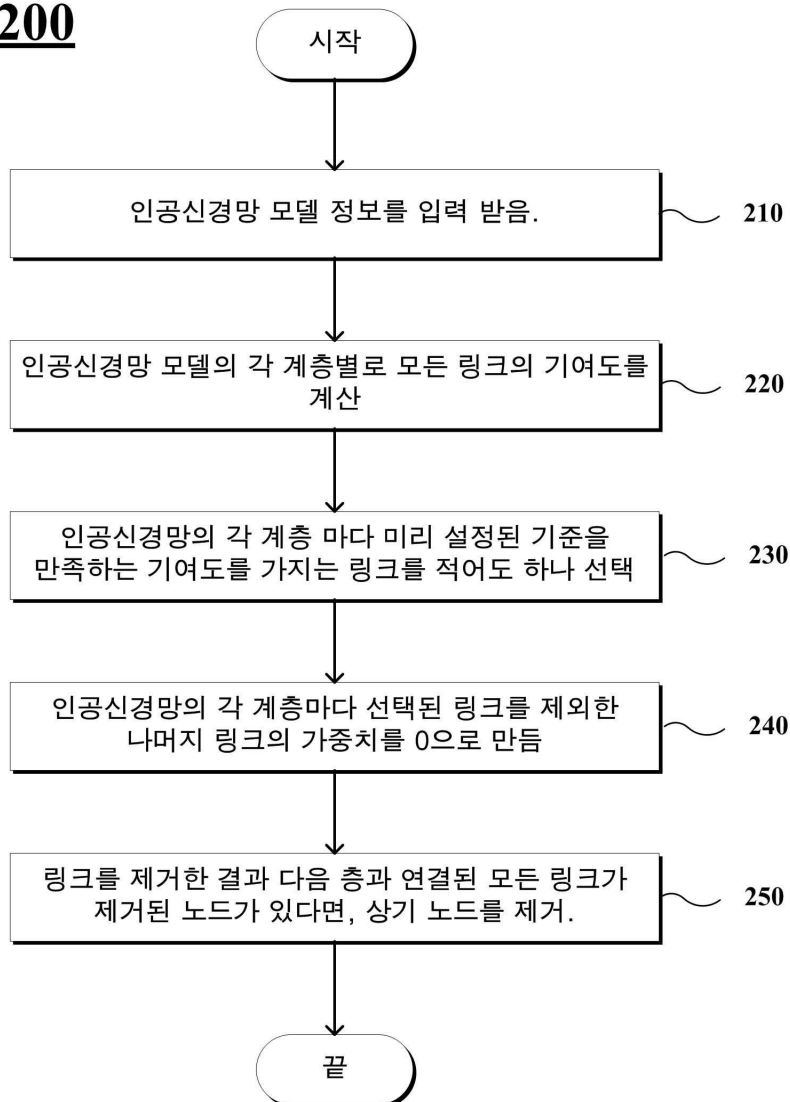
도면

도면1

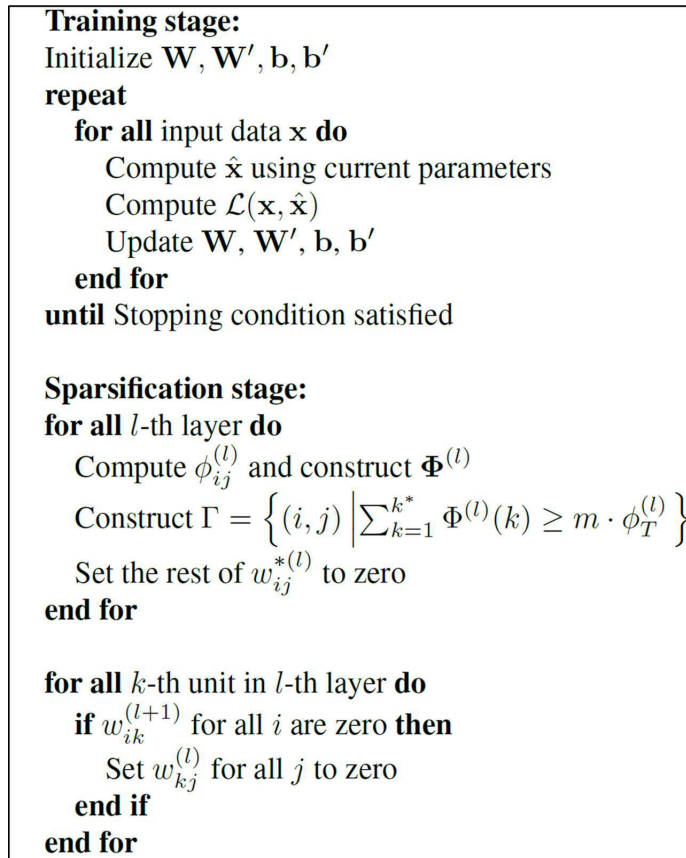


도면2

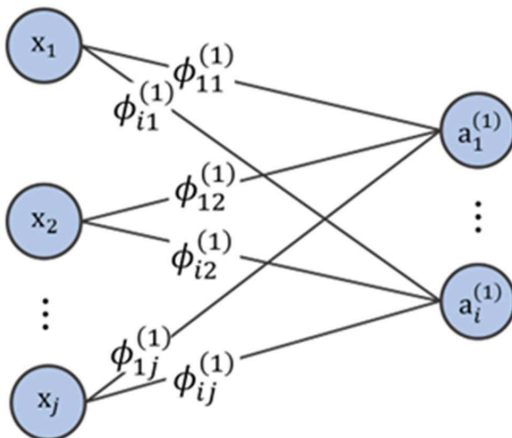
200



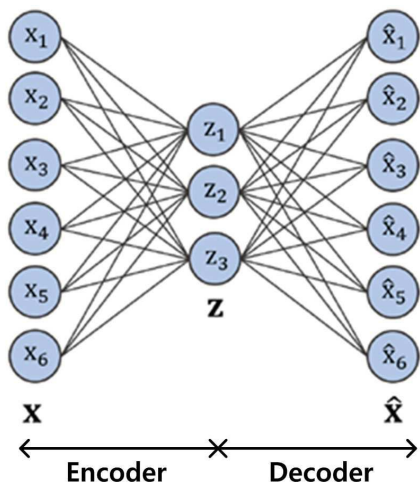
도면3



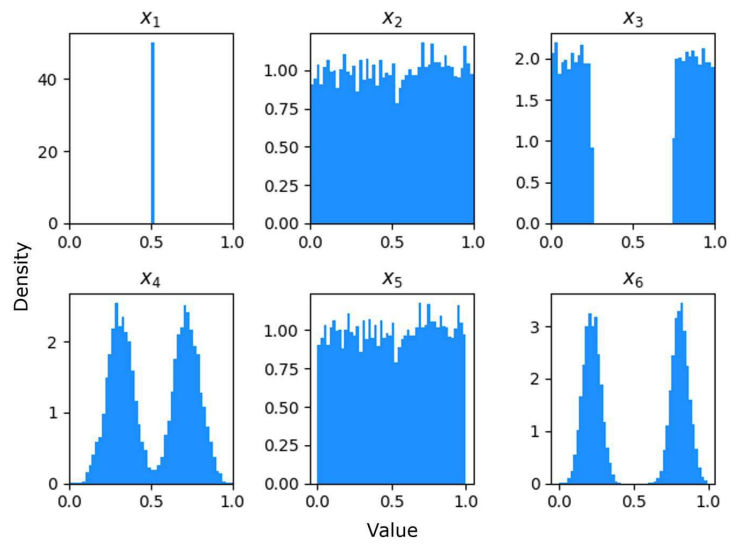
도면4



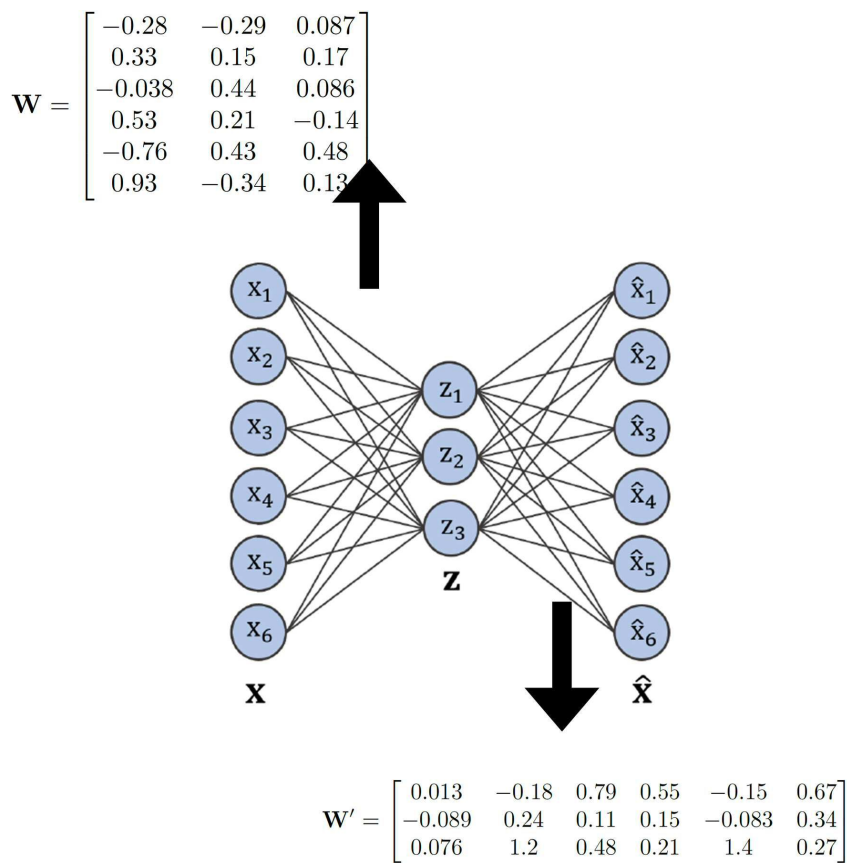
도면5



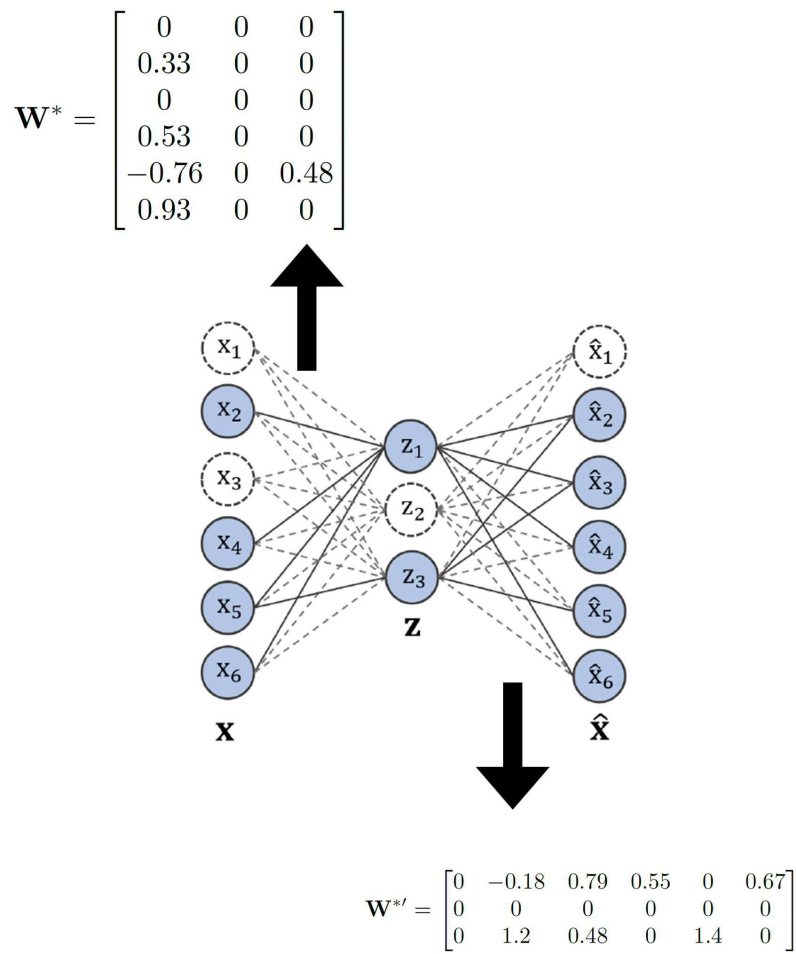
도면6



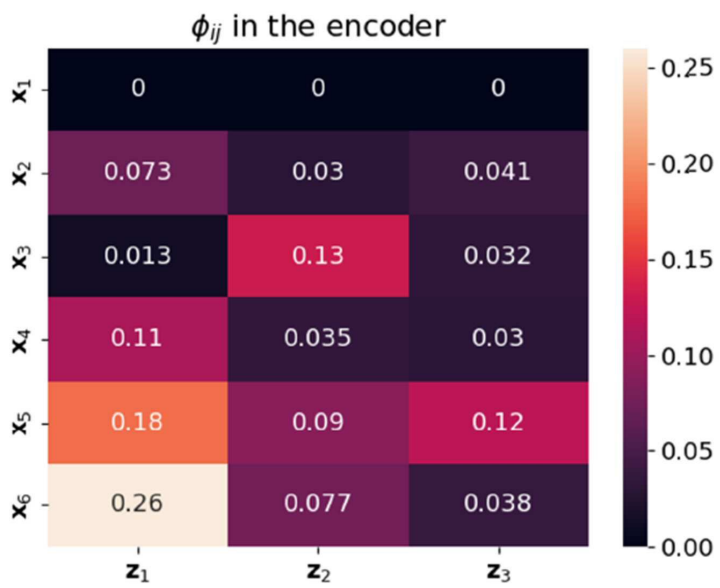
도면7



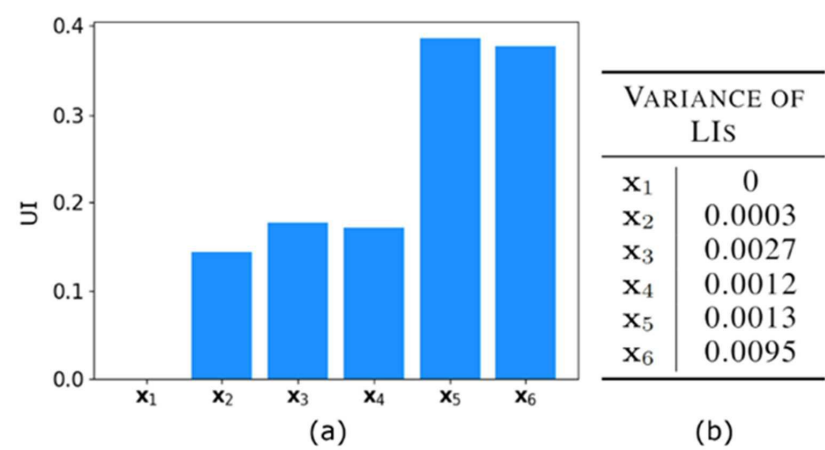
도면8



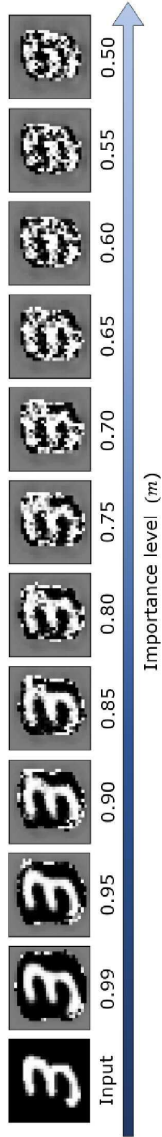
도면9



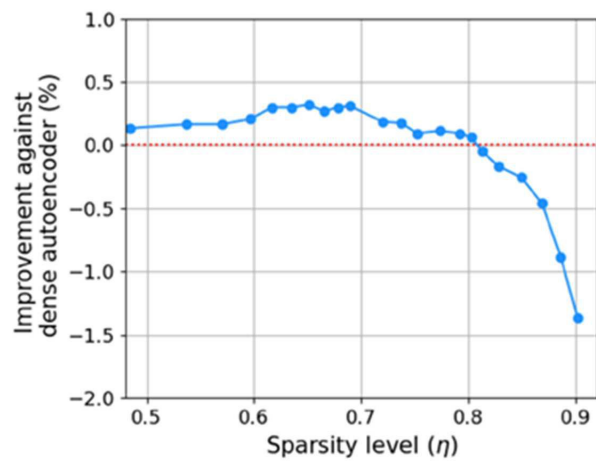
도면10



도면11



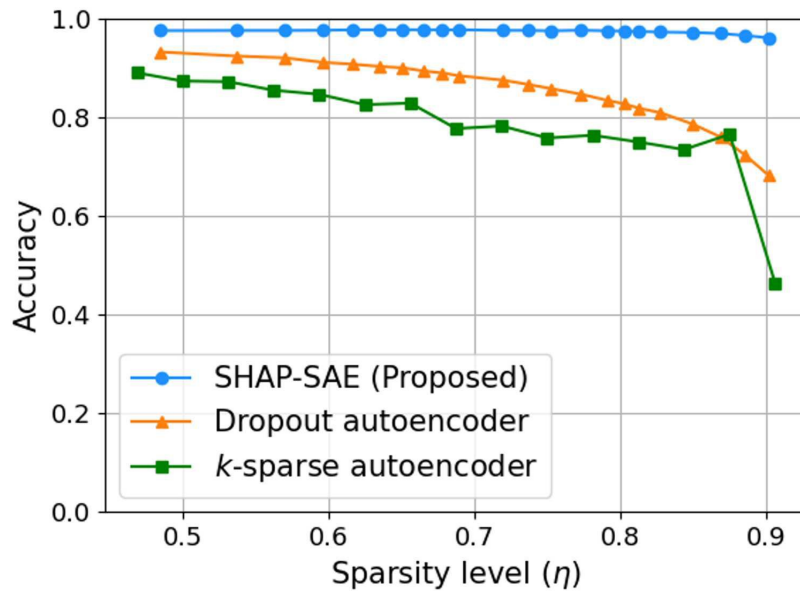
도면12



도면13

m	Output	η	Accuracy (%)	Performance improvement against dense autoencoder (%)
0.99		0.49	97.47	0.13
0.98		0.54	97.50	0.16
0.97		0.57	97.50	0.16
0.96		0.60	97.54	0.21
0.95		0.62	97.63	0.30
0.94		0.64	97.63	0.30
0.93		0.65	97.65	0.32
0.92		0.67	97.60	0.27
0.91		0.68	97.63	0.30
0.90		0.69	97.64	0.31
0.85		0.74	97.43	0.17
0.80		0.77	97.45	0.11
0.75		0.80	97.40	0.06
0.70		0.83	97.18	-0.16
0.65		0.85	97.09	-0.26
0.60		0.87	96.89	-0.46
0.55		0.89	96.48	-0.88
0.50		0.90	96.01	-1.37

도면14



도면15

SPARSITY LEVEL (η)	METHOD	ERROR RATE
0	DENSE AUTOENCODER	2.66%
0.54	<i>k</i> -SPARSE AUTOENCODER (<i>k</i> =15)	12.87%
	DROPOUT AUTOENCODER	7.87%
	SHAP-SAE (<i>m</i>=0.98)	2.50 %
0.69	<i>k</i> -SPARSE AUTOENCODER (<i>k</i> =10)	22.38%
	DROPOUT AUTOENCODER	11.78%
	SHAP-SAE (<i>m</i>=0.90)	2.36 %
[0.77, 0.78]	<i>k</i> -SPARSE AUTOENCODER (<i>k</i> =7)	26.63%
	DROPOUT AUTOENCODER	23.74%
	SHAP-SAE (<i>m</i>=0.80)	2.55 %
[0.84, 0.85]	<i>k</i> -SPARSE AUTOENCODER (<i>k</i> =5)	26.63%
	DROPOUT AUTOENCODER	22.14%
	SHAP-SAE (<i>m</i>=0.65)	2.91 %
0.90	<i>k</i> -SPARSE AUTOENCODER (<i>k</i> =3)	53.80%
	DROPOUT AUTOENCODER	30.88%
	SHAP-SAE (<i>m</i>=0.50)	3.99 %

도면16

