



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0018331
(43) 공개일자 2025년02월05일

- | | |
|---|---|
| <p>(51) 국제특허분류(Int. Cl.)
G06N 3/0495 (2023.01) G06N 3/0475 (2023.01)
G06N 3/08 (2023.01) G06N 3/096 (2023.01)</p> <p>(52) CPC특허분류
G06N 3/0495 (2023.01)
G06N 3/0475 (2023.01)</p> <p>(21) 출원번호 10-2023-0099154
(22) 출원일자 2023년07월28일
심사청구일자 2023년07월28일</p> | <p>(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50</p> <p>(72) 발명자
유재준
울산광역시 울주군 언양읍 유니스트길 50
여상엽
울산광역시 울주군 언양읍 유니스트길 50
(뒷면에 계속)</p> <p>(74) 대리인
특허법인비엘티</p> |
|---|---|

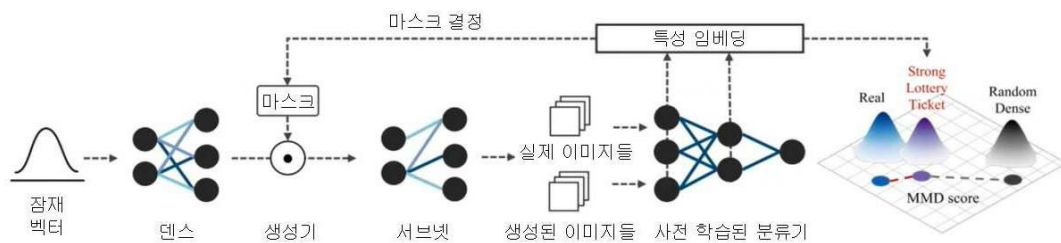
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 서버 네트워크를 통한 인공지능 기반 생성모델 경량화 장치 및 방법

(57) 요약

본 개시는 서버 네트워크를 통한 인공지능 기반 생성모델 경량화 장치 및 방법에 관한 것으로, 외부 장치와 통신을 수행하는 통신모듈; 인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 적어도 하나의 프로세스를 저장하는 저장모듈; 및 상기 적어도 하나의 프로세스를 기반으로 상기 인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 동작을 수행하는 제어모듈을 포함하며, 상기 제어모듈은, 엣지 팝업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하고, 임의의 서버 네트워크를 찾은 후, 각각의 순방향 경로에서 할당된 스코어를 정렬하여 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트 할 수 있다.

대표도



(52) CPC특허분류

G06N 3/084 (2023.01)

G06N 3/096 (2023.01)

(72) 발명자

장유진

울산광역시 울주군 언양읍 유니스트길 50

손지용

서울특별시 서대문구 연세로 50 연세대학교 대우관
534호

한동윤

서울시 성동구 행당로

이 발명을 지원한 국가연구개발사업	
과제고유번호	1711189018
과제번호	2022R1C1C1008496
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	강력하고 해석 가능한 딥러닝 생성 모델을 위한 신호처리 기반 분석 도구 및 학습 알고리즘 개발: 일반, 바이오, 의료 영상으로의 응용
기 여 율	40/100
과제수행기관명	울산과학기술원
연구기간	2023.03.01 ~ 2024.02.29
이 발명을 지원한 국가연구개발사업	
과제고유번호	1711193127
과제번호	2022-0-00959-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	(2세부) 의사결정 지원을 위한 퓨샷 학습 기반 시각 및 언어에 대한 인과관계 추론 기술개발
기 여 율	40/100
과제수행기관명	서울대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업	
과제고유번호	1711193983
과제번호	2020-0-01336-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(울산과학기술원)
기 여 율	5/100
과제수행기관명	울산과학기술원
연구기간	2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업	
과제고유번호	1711193622
과제번호	2021-0-02068-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능 혁신 허브 연구 개발
기 여 율	5/100
과제수행기관명	고려대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업	
과제고유번호	1711193397
과제번호	2022-0-00264-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	지식기반 심층논리 신경망을 활용한 통합적 비디오 해석과 생성 연구
기 여 율	10/100
과제수행기관명	한양대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

외부 장치와 통신을 수행하는 통신모듈;

인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 적어도 하나의 프로세스를 저장하는 저장모듈; 및

상기 적어도 하나의 프로세스를 기반으로 상기 인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 동작을 수행하는 제어모듈을 포함하며,

상기 제어모듈은,

엣지 팝업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하고, 임의의 서브 네트워크를 찾은 후, 각각의 순방향 경로에서 할당된 스코어를 정렬하여 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트 하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 2

제1항에 있어서,

상기 제어모듈은,

상기 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 다른 웨이트를 0으로 만들고, 역방향 경로에서는 상기 서브 네트워크의 로스를 산출하여 역전파를 활용하도록 하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 3

제1항에 있어서,

상기 제어모듈은,

상기 서브 네트워크의 로스를 산출할 시, 상기 서브 네트워크를 통해 생성된 이미지와 실제 이미지를 임베딩 스캐이스로 보내 최대 평균 불일치(Maximum Mean Discrepancy, MMD) 스코어를 산출하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 4

제3항에 있어서,

상기 제어모듈은,

두 개의 샘플 셋으로서 실제 샘플 및 가짜 샘플에 대해 모든 차수의 모멘트를 매칭함으로써 상기 최대 평균 불일치를 산출하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 5

제4항에 있어서,

상기 최대 평균 불일치는,

하기 <수학식 1>을 기초로 산출되는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

[수학식 1]

$$L_{MMD} = \left\| \frac{1}{N} \sum_{i=1}^N \phi(r_i) - \frac{1}{M} \sum_{j=1}^M \phi(f_j) \right\|^2 \dots$$

여기서, r_i 는 진짜 샘플을 나타내고, f_j 는 가짜 샘플을 나타내는 것임.

청구항 6

제1항에 있어서,

상기 제어모듈은,

상기 모멘트 매칭을 위해 사전 학습된 VGG 네트워크를 커널로 하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 7

제6항에 있어서,

상기 제어모듈은,

상기 스코어를 업데이트 하는 동작을 반복적으로 수행함으로써, SLTs(Strong Lottery Tickets)를 찾아내는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치.

청구항 8

장치에 의해 수행되는, 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법에 있어서,

엣지 팝업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하는 단계;

임의의 서브 네트워크를 찾는 단계;

각각의 순방향 경로에서 할당된 스코어를 정렬하는 단계; 및

기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트 하는 단계를 포함하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법.

청구항 9

제8항에 있어서,

상기 업데이트 하는 단계는,

상기 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 다른 웨이트를 0으로 만들고, 역방향 경로에서는 상기 서브 네트워크의 로스를 산출하여 역전파를 활용하도록 하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법.

청구항 10

제1항에 있어서,

상기 업데이트 하는 단계는,

상기 서브 네트워크의 로스를 산출할 시, 상기 서브 네트워크를 통해 생성된 이미지와 실제 이미지를 임베딩 스케이스로 보내 최대 평균 불일치(Maximum Mean Discrepancy, MMD) 스코어를 산출하는 것을 특징으로 하는,

서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법.

발명의 설명

기술 분야

[0001] 개시는 인공지능 기반 생성모델 경량화 장치 및 방법에 대한 것으로, 보다 상세하게는, 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치 및 방법에 관한 것이다.

배경 기술

[0002] 최첨단 생성모델은 더 나은 성능을 위해 매우 크고 복잡한 구조를 사용하는 경향이 있다.

[0003] 그러나, 대형 모델의 한가지 단점은 트레이닝을 위한 높은 컴퓨팅 비용으로 인해 모바일 환경과 같은 엣지 디바이스에 적용하는데에는 제한이 있다는 것이다.

[0004] 그렇기 때문에 자연스럽게 새로운 경량 아키텍처 또는 제너레이티브 모델링의 새로운 압축 방법을 설계할 필요가 있다.

[0005] 일반적으로, 인공지능 기반의 생성모델을 경량화하기 위해 훈련(train) - 가지치기(prune) - 재훈련(retraining)의 과정을 거치도록 하고 있으나, 불안정성으로 인해 추가적인 네트워크가 요구되거나 복잡한 학습 과정으로 인하여 추가 비용이 필요한 문제점이 있다.

[0006] 따라서, 인공지능 기반의 생성모델을 안정적으로 압축할 수 있고, 성능의 저하 없이 경량화할 수 있도록 하는 기술이 개발될 필요가 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국공개특허공보 제10-2023-0073608호 (공개일: 2023년 05월 26일)

발명의 내용

해결하려는 과제

[0008] 본 개시는 서브 네트워크 중 학습된 생성모델과 같은 네트워크를 안정적으로 찾기 위한 서브 네트워크(strong lottery tickets, SLTs) 알고리즘을 이용하여 생성모델을 안정적으로 압축할 수 있고 성능의 저하 없이 경량화할 수 있도록 하는 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치 및 방법을 제공함에 있다.

[0009] 본 개시가 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0010] 상술한 기술적 과제를 달성하기 위한 본 개시의 일 측면에 따른 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치는, 외부 장치와 통신을 수행하는 통신모듈; 인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 적어도 하나의 프로세스를 저장하는 저장모듈; 및 상기 적어도 하나의 프로세스를 기반으로 상기 인공지능을 기반으로 생성모델을 경량화하여 제공하기 위한 동작을 수행하는 제어모듈을 포함하며, 상기 제어모듈은, 엣지 팝업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하고, 임의의 서브 네트워크를 찾은 후, 각각의 순방향 경로에서 할당된 스코어를 정렬하여 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트할 수 있다.

[0011] 한편, 본 개시의 일 측면에 따른 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법은, 엣지 팝업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하는 단계; 임의의 서브 네트워크를 찾는 단계; 각각의 순방향 경로에서 할당된 스코어를 정렬하는 단계; 및 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트 하는 단계를 포함할 수 있다.

[0012] 이 외에도, 본 개시를 구현하기 위한 방법을 실행하기 위한 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램이 더 제공될 수 있다.

[0013] 이 외에도, 본 개시를 구현하기 위한 방법을 실행하기 위한 컴퓨터 프로그램을 기록하는 컴퓨터 판독 가능한 기록 매체가 더 제공될 수 있다.

발명의 효과

[0014] 본 개시의 전술한 과제 해결 수단에 의하면, 서브 네트워크 중 학습된 생성모델과 같은 네트워크를 안정적으로 찾기 위한 서브 네트워크(strong lottery tickets, SLTs)알고리즘을 이용하여 생성모델을 안정적으로 압축할 수 있고 성능의 저하 없이 경량화할 수 있도록 한다.

[0015] 본 개시의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0016] 도 1은 일반적인 인공지능 기반 생성모델 경량화를 위한 동작을 설명하기 위한 인공지능 신경망의 블록도

도 2는 본 개시의 일 실시예에 따라 서브 네트워크를 통한 인공지능 기반 생성모델 경량화를 위한 일련의 동작 절차를 나타내는 도면

도 3은 본 개시에 따른 다른 서브 네트워크를 통한 인공지능 기반 생성모델 경량화를 위해 STLs를 찾는 동작을 설명하기 위한 도면

도 4는 본 개시에 따라 가중치에 적절한 스코어를 할당하는 동작을 구체적으로 설명하기 위한 도면

도 5는 본 개시에 따라 모멘트 매칭을 통해 안정적인 스코어를 모델링하는 동작을 구체적으로 설명하기 위한 도면

도 6은 본 개시의 일 실시예에 따라 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 장치의 구조를 나타내는 도면

도 7은 본 개시의 일 실시예에 따라 서브 네트워크를 통한 인공지능 기반 생성모델 경량화 방법을 나타내는 도면

도 8은 본 개시에 따라 생성모델에서 SLTs가 존재하는지 여부를 확인하기 위해 진행한 실험 결과의 일 예시를 나타내는 도면

발명을 실시하기 위한 구체적인 내용

[0017] 본 개시의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 개시는 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 개시가 완전하도록 하고, 본 개시가 속하는 기술 분야의 통상의 기술자에게 본 개시의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 개시는 청구항의 범주에 의해 정의될 뿐이다.

[0018] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 개시를 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 개시의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0019] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 개시가 속하는 기술 분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지

않는다.

- [0020] 본 개시 전체에 걸쳐 동일 참조 부호는 동일 구성요소를 지칭한다. 본 개시가 실시예들의 모든 요소들을 설명하는 것은 아니며, 본 개시가 속하는 기술분야에서 일반적인 내용 또는 실시예들 간에 중복되는 내용은 생략한다. 명세서에서 사용되는 "부" 또는 "모듈"이라는 용어는 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소를 의미하며, "부" 또는 "모듈"은 어떤 역할들을 수행한다. 그렇지만 "부" 또는 "모듈"은 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부" 또는 "모듈"은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 "부" 또는 "모듈"은 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 "부" 또는 "모듈"들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 "부" 또는 "모듈"들로 결합되거나 추가적인 구성요소들과 "부" 또는 "모듈"들로 더 분리될 수 있다.
- [0021] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 직접적으로 연결되어 있는 경우뿐 아니라, 간접적으로 연결되어 있는 경우를 포함하고, 간접적인 연결은 무선 통신망을 통해 연결되는 것을 포함한다.
- [0022] 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.
- [0023] 명세서 전체에서, 어떤 부재가 다른 부재 "상에" 위치하고 있다고 할 때, 이는 어떤 부재가 다른 부재에 접해 있는 경우뿐 아니라 두 부재 사이에 또 다른 부재가 존재하는 경우도 포함한다.
- [0024] 제1, 제2 등의 용어는 하나의 구성요소를 다른 구성요소로부터 구별하기 위해 사용되는 것으로, 구성요소가 전술된 용어들에 의해 제한되는 것은 아니다.
- [0025] 단수의 표현은 문맥상 명백하게 예외가 있지 않는 한, 복수의 표현을 포함한다.
- [0026] 각 단계들에 있어 식별부호는 설명의 편의를 위하여 사용되는 것으로 식별부호는 각 단계들의 순서를 설명하는 것이 아니며, 각 단계들은 문맥상 명백하게 특정 순서를 기재하지 않는 이상 명기된 순서와 다르게 실시될 수 있다.
- [0027] 이하 설명에서 사용되는 용어에 대하여 정의하면 하기와 같다.
- [0028] 본 명세서에서 인공지능 기반 사전 학습모델은 딥러닝 기반의 예측 모델이며, 이들 각각을 통해 수술 전 또는 수술 중에 해당 환자의 수술 부위에 대한 재파열 확률을 예측하여 예측정보를 생성할 수 있다. 이때, 딥러닝 방식은 한정하지 않으며, 상황(필요)에 따라 적어도 하나의 방식이 적용될 수 있다. 이때, 인공지능 알고리즘의 일 예로서 RNN(Recurrent Neural Network) 또는 트랜스포머(transformer) 등이 적용될 수 있으나, 이것으로 한정하는 것은 아니며, 다른 인공지능 알고리즘이 적용될 수도 있다.
- [0029] 본 명세서에서 '경량화 장치(100)'로 한정하여 설명하였으나, 이는 서버 네트워크를 통해 인공지능 기반 생성모델을 경량화하여 제공하기 위한 장치로서, 연산처리를 수행할 수 있는 다양한 장치들을 모두 포함할 수 있다. 즉, 경량화 장치(100)는 서버, 컴퓨터, 서버 및/또는 휴대용 단말기를 더 포함하거나, 또는 그 중 어느 하나의 형태가 될 수 있으며, 이를 한정하지 않는다.
- [0030] 여기에서, 상기 컴퓨터는 예를 들어, 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크톱(desktop), 랩톱(laptop), 태블릿 PC, 슬레이트 PC 등을 포함할 수 있다.
- [0031] 상기 서버는 외부 장치와 통신을 수행하여 정보를 처리하는 서버로써, 애플리케이션 서버, 컴퓨팅 서버, 데이터베이스 서버, 파일 서버, 게임 서버, 메일 서버, 프록시 서버 및 웹 서버 등을 포함할 수 있다.
- [0032] 상기 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, PCS(Personal Communication System), GSM(Global System for Mobile communications), PDC(Personal Digital Cellular), PHS(Personal Handyphone System), PDA(Personal Digital Assistant), IMT(International Mobile Telecommunication)-2000, CDMA(Code Division Multiple Access)-2000, W-CDMA(W-Code Division Multiple Access), WiBro(Wireless Broadband Internet) 단말, 스마트폰(Smart Phone) 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치와 시계, 반지, 팔찌, 발찌, 목걸이, 안경, 콘택트 렌즈, 또는 머리 착용형

장치(head-mounted-device(HMD) 등과 같은 웨어러블 장치를 포함할 수 있다.

- [0033] 이하 첨부된 도면들을 참고하여 본 개시의 작용 원리 및 실시예들에 대해 설명한다.
- [0034] 도 1은 일반적인 인공지능 기반 생성모델 경량화를 위한 동작을 설명하기 위한 인공지능 신경망의 블록도이다.
- [0035] 도 1에 도시된 바와 같이, 랜덤 네트워크, 즉, 학습이 되지 않은 모델은 좋지 않은 성능을 보여주므로, 이를 학습하여 높은 성능의 네트워크(학습된 모델)를 얻는다. 이후, 그 학습된 모델을 기준으로 특정 기준을 제안해서 가중치를 없애는 등의 방식을 통해 가지치기(pruning)를 진행하는데, 이 경우 네트워크의 구조가 바뀌기 때문에 네트워크의 성능이 감소하게 된다. 이를 다시 복원해주기 위해 재학습을 수행하게 되는데, 결론적으로 낮은 성능의 네트워크가 생성된다.
- [0036] 즉, 그 과정에서 학습이 여러 번 진행됨에 따라 성능이 감소하는 문제가 발생할 수밖에 없다.
- [0037] 앞서 살펴본 바와 같이, 기존의 가지치기 알고리즘은 과도한 가중치 훈련 비용, 성능 저하, 제한된 일반화 가능성, 복잡한 훈련 등의 문제를 안고 있는 것을 알 수 있다.
- [0038] 이러한 문제점을 해결하기 위해, 본 개시는 가중치 업데이트 없이도 생성모델에서 우수한 생성 성능을 달성하는 서브 네트워크인 SLTs를 찾도록 하기 위한 것으로, 이때, 모멘트 매칭 스코어(Moment Matching Score)를 통해 SLTs를 찾도록 한다. 사전 학습된 분류기의 성능을 활용하여 랜덤 초기화된 가중치에 스코어를 할당하고, 스파스 마스크(sparse mask)를 찾아내어 서브 네트워크가 훈련된 텐스 네트워크(dense network/dense generator)와 비슷하거나 더 나은 성능을 발휘하도록 한다.
- [0039] 이 SLTs는 초기화 시(즉, 가중치 업데이트가 없는) 서브 네트워크로, 가중치가 학습된 dense counterpart과 비슷하거나 더 나은 성능을 보인다. 여기서는, 변별적 모델에서 SLTs를 찾는 가장 초기의 방법으로 엡지 팝업 알고리즘(edge-popup algorithm)을 이용한다. 이 엡지 팝업 알고리즘은 각 가중치의 중요도를 스코어화할 수 있다는 아이디어를 기반으로 서브 네트워크 마스크를 선택한다. 이러한 스코어가 할당되면 원하는 목표 최소도에 따라 높은 스코어의 가중치를 유지하기만 하면 된다.
- [0040] 이 엡지 팝업 알고리즘의 성능은 가지치기 기준(pruning criteria)이 되는 업데이트 된 스코어에 따라 크게 좌우되기 때문에, 생성모델의 가지치기를 위해서는 적절한 스코어 함수를 사용하는 것이 필수적이다. 고품질 생성기 훈련에 일반적으로 사용되는 기준인 적대적 손실을 쉽게 생각할 수 있지만, 이는 매우 불안정하고 적절한 스코어를 찾는 데 방해가 된다.
- [0041] 여기서, 본 개시는 최대 평균 불일치(Maximum Mean Discrepancy, MMD)로 알려진 통계적 가설 검증 기법을 활용하도록 한다. 이 기법은 고정된 사전 훈련된 ConvNet에서 추출한 특징을 사용하여 간단한 순간 매칭 스코어를 도출한다.
- [0042] 즉, 본 개시는 엡지 팝업 알고리즘 및 모멘트 매칭 스코어를 결합하여 매우 희박한 체제에서 우수한 생성 성능을 가진 서브 네트워크를 찾는 안정적인 알고리즘을 제공한다. 이는 가중치 업데이트가 필요하지 않기 때문에 훈련과 가지치기 과정 사이의 균형을 맞추는 까다로운 문제를 피할 수 있다. 또한, 모멘트 매칭 스코어의 안정된 특성으로 인하여 부가적인 기능 없이도 SLTs를 찾을 수 있다.
- [0043] 무작위로 초기화된 가중치 $\theta \in \mathbb{R}^d$ 를 가진 신경망($z; \theta$)을 생성한다. Rm 다음으로, SLTs를 찾는 것을 목표로 한다. 그 다음 SLTs를 찾는 것을 목표로 할 때, 마스크 $m \in \{0, 1\}$ 는 가지치기된 신경망 $G(z; \theta \odot m)$ 가 생성 태스크에서 잘 수행됨을 만족하는 것이다.
- [0044] 도 2는 본 개시의 일 실시예에 따라 서브 네트워크를 통한 인공지능 기반 생성모델 경량화를 위한 일련의 동작 절차를 나타내는 도면이다.
- [0045] 도 2를 참조하면, 본 개시의 일 실시예에 따른 경량화 장치(100)는 텐스 네트워크가 있을 때, 가중치에 임의의 스코어를 할당하고, 임의의 서버 네트워크를 찾는다.
- [0046] 그 다음으로, 생성된 이미지(generated images)와 실제 이미지(real images)를 임베딩 스페이스(Embedding space)로 보내 MMD 스코어를 산출한 후, 앞서 할당된 스코어를 업데이트하여 마스크를 업데이트 해주는 방식으로 그 서버 네트워크를 계속적으로 업데이트 해나간다.
- [0047] 결론적으로, 가중치 업데이트 없이도 생성모델에서 SLTs를 찾아낼 수 있게 된다.
- [0048] 도 3은 본 개시에 따른 다른 서버 네트워크를 통한 인공지능 기반 생성모델 경량화를 위해 SLTs를 찾는 동작을

설명하기 위한 도면이다.

- [0049] 앞서 설명한 바와 같이, 본 개시는 무작위화 된 판별 네트워크에서 SLTs를 찾는 가장 초기의 방법으로 엣지 팝업 알고리즘을 이용하는데, 이 엣지 팝업 알고리즘에서는 랜덤 초기화된 가중치(ω)가 있을 때, 그 가중치 중에 어떤 가중치가 중요한지를 찾아야하기 때문에, 얼마나 중요한지에 대한 스코어 s 를 할당한다. 구체적으로, 각 가중치 θ_i 에 대해 무작위 스코어 s_i 를 할당한다($\theta = [\theta_1, \dots, \theta_d]$). 이때, 가중치를 $k\%$ 로 유지한다고 가정한다. 그런 다음 각 순방향 경로에서 각 레이어에서 스코어 s_i 를 정렬라고 해당 레이어 내에서 $|s_i|$ 가 상위 $k\%$ 에 속하면 $m_i = 1$ 을 할당하고, 그렇지 않으면 $m_i = 0$ 을 할당한다. 역방향 경로에 대해, 네트워크의 손실을 계산하고 역전파를 사용하여 스코어 s_i 를 업데이트 한다. 여기서는, 직선 추정기(straight-through estimator)를 사용하여 s_i 를 m_i 로 매핑하는 지표 함수를 처리한다.
- [0050] 도 4는 본 개시에 따라 가중치에 적절한 스코어를 할당하는 동작을 구체적으로 설명하기 위한 도면이다.
- [0051] 한편, 적절한 스코어를 배정하기 위해서 역전파(back-propagation)를 이용하여 랜덤 초기화된 스코어 s 를 업데이트 한다. 네트워크의 서브 네트워크를 찾아서 그 출력(output)을 구하고 이것을 역전파를 통해서 전체 스코어를 업데이트해준다. 즉, 스코어를 업데이트 함으로써 적절한 스코어를 찾을 수 있을 것을 기대할 수 있다.
- [0052] 도 5는 본 개시에 따라 모멘트 매칭을 통해 안정적인 스코어를 모델링하는 동작을 구체적으로 설명하기 위한 도면이다.
- [0053] 한편, 생성모델을 가지치기하려면 판별 모델에 사용되는 교차 엔트로피 손실 대신 적절한 스코어 업데이트 함수가 필요하다. 이를 위해, 생성모델의 학습에 안정적인 최적화를 제공하는 것으로 알려진 최대 평균 불일치(Maximum Mean Discrepancy, MMD)를 이용한다.
- [0054] 주어진 실제 샘플 $\{r_i\}_{i=1}^N$ 과 가짜 샘플 $\{f_j\}_{j=1}^M$ 의 두 세트가 주어진 경우, MMD 손실 L_{MMD} 를 최소화하는 것은 모델 분포의 모든 순간을 경험적 데이터 분포와 일치시키는 것으로 해석할 수 있다. 이 L_{MMD} 는 하기 <수학식 1>을 기초로 산출될 수 있다.

수학식 1

$$\mathcal{L}_{\text{MMD}} = \left\| \frac{1}{N} \sum_{i=1}^N \phi(r_i) - \frac{1}{M} \sum_{j=1}^M \phi(f_j) \right\|^2, \quad (1)$$

$$\begin{aligned} \mathcal{L}_{\text{MMD}} = & \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N \psi(r_i, r_{i'}) - \frac{2}{NM} \sum_{i=1}^N \sum_{j=1}^M \psi(r_i, f_j) \\ & + \frac{1}{M^2} \sum_{j=1}^M \sum_{j'=1}^M \psi(f_j, f_{j'}), \end{aligned} \quad (2)$$

$$\mathcal{L}_{\text{MMD}} = \sum_{j=1}^L \left\| \mu_r^j - \mu_f^j \right\|^2 + \left\| \sigma_r^j - \sigma_f^j \right\|^2. \quad (3)$$

- [0055]
- [0056] 먼저, 식 (1)에서 $\phi(\cdot)$ 는 교차 모멘트를 일치시키는 함수를 나타낸다. 이상적으로 $\phi(\cdot)$ 는 무한 차수로 계산해야 한다. MMD를 효율적으로 계산하기 위해서는 이 식 (1)을 커널 트릭을 통해 호출한다.
- [0057] 또한, 사전 훈련된 VGG 네트워크를 고정된 커널 ψ 로 사용하고, VGG 임베딩 공간에서 실제 샘플 및 가짜 샘플 특징의 평균 μ 와 공분산 σ 를 일치시킨다. 이는 강력한 커널을 활용할수록 더 좋은 성능을 낼 수 있기 때문에,

학습된 네트워크(Pretrained fixed feature extractor)가 좋은 커널로서의 역할을 할 수 있기 있으므로 이 VGG 네트워크를 사용 그 고정된 커널 ψ 를 사용한 것이다.

[0058] 이로써, 이 VGG 네트워크를 통해서 임베딩된 리얼 데이터와 페이크 데이터를 통해 MMD 로스를 구해서 스코어를 업데이트해준다.

[0059] 한편, I_v , w_{uv} , σ , 및 α 를 각각 노드 v 의 입력, 노드 u 와 노드 v 의 네트워크 파라미터, 활성화 함수 및 학습률로 정의한다. 시간 단계 t 에서 스코어의 변화량은 하기 <수학식 2>와 같이 나타낼 수 있다.

수학식 2

$$S_{t+1,uv} = S_{t,uv} - \alpha \frac{\partial \mathcal{L}_{MMD}}{\partial I_v} w_{uv} \sigma(I_u). \quad (4)$$

[0060]

[0061] 이러한 방법은 학습 가중치가 아니라 중요도가 낮은 노드를 찾기 위해 MMD 손실을 사용한다는 점은 주목할 가치가 있다.

[0062] 도 6은 본 개시의 일 실시예에 따라 서버 네트워크를 통한 인공지능 기반 생성모델 경량화 장치의 구조를 나타내는 도면이다.

[0063] 도 6을 참조하면, 본 개시의 일 실시예에 따른 서버 네트워크를 통한 인공지능 기반 생성모델 경량화 장치(이하, '경량화 장치'라 칭함)(100)는 통신모듈(110), 저장모듈(130) 및 제어모듈(150)을 포함하여 구성될 수 있다.

[0064] 통신모듈(110)은 적어도 하나의 장치/단말과 적어도 하나의 정보 또는 데이터를 송수신한다. 여기서, 적어도 하나의 장치/단말은 인공지능 기반 생성모델을 경량화한 경량화 모델을 제공받는 장치/단말이거나, 그 생성모델을 경량화 하기 위해 필요한 각종 데이터/정보를 제공하는 장치/단말일 수 있으며, 그 종류 및 형태를 한정하지 않는다.

[0065] 또한, 이 통신모듈(110)은 그 외 다른 장치들과의 통신을 수행할 수도 있는 것으로, 무선 인터넷 기술들에 따른 통신망에서 무선 신호를 송수신하도록 한다.

[0066] 무선 인터넷 기술로는, 예를 들어 WLAN(Wireless LAN), Wi-Fi(Wireless-Fidelity), Wi-Fi(Wireless Fidelity) Direct, DLNA(Digital Living Network Alliance), WiBro(Wireless Broadband), WiMAX(World Interoperability for Microwave Access), HSDPA(High Speed Downlink Packet Access), HSUPA(High Speed Uplink Packet Access), LTE(Long Term Evolution), LTE-A(Long Term Evolution-Advanced) 등이 있으며, 경량화 장치(100)는 앞에서 나열되지 않은 인터넷 기술까지 포함한 범위에서 적어도 하나의 무선 인터넷 기술에 따라 데이터를 송수신하게 된다.

[0067] 근거리 통신(Short range communication)을 위한 것으로서, 블루투스(Bluetooth™), RFID(Radio Frequency Identification), 적외선 통신(Infrared Data Association; IrDA), UWB(Ultra Wideband), ZigBee, NFC(Near Field Communication), Wi-Fi(Wireless-Fidelity), Wi-Fi Direct, Wireless USB(Wireless Universal Serial Bus) 기술 중 적어도 하나를 이용하여, 근거리 통신을 지원할 수 있다. 이러한, 근거리 무선 통신망(Wireless Area Networks)을 경량화 장치(100)와, 적어도 하나의 사용자 단말(미도시) 간 무선 통신을 지원할 수 있다. 이때, 근거리 무선 통신망은 근거리 무선 개인 통신망(Wireless Personal Area Networks)일 수 있다.

[0068] 저장모듈(130)은 서버 네트워크를 통한 인공지능 기반 생성모델 경량화하여 제공하기 위한 적어도 하나의 프로세스(알고리즘)는 또는 그 프로세스를 재현한 프로그램에 대한 데이터를 저장할 수 있다. 뿐만 아니라, 저장모듈(130)은 그 외 다른 동작을 수행하기 위한 프로세스들을 더 저장할 수 있으며, 이를 한정하지 않는다.

[0069] 저장모듈(130)은 서버 네트워크를 통한 인공지능 기반 생성모델 경량화하여 제공하기 위해 필요한 각종 정보/데이터는 물론, 경량화 장치(100)의 다양한 기능을 지원하는 그 외 다양한 데이터들을 저장할 수 있다. 저장모듈(130)은 경량화 장치(100)에서 구동되는 다수의 응용 프로그램(application program 또는 애플리케이션(application), 경량화 장치(100)의 동작을 위한 데이터들, 명령어들을 저장할 수 있다. 이러한 응용 프로그램 중 적어도 일부는, 무선 통신을 통해 외부 서버로부터 다운로드 될 수 있다. 한편, 응용 프로그램은, 저장모듈

(130)에 구비된 적어도 하나의 메모리에 저장되고, 경량화 장치(100) 상에 설치되어, 제어모듈(150)을 통해 저장모듈(130)에 저장된 적어도 하나의 프로세서에 의하여 동작(또는 기능)을 수행하도록 구동될 수 있다.

[0070] 한편, 적어도 하나의 메모리는 플래시 메모리 타입(flash memory type), 하드 디스크 타입(hard disk type), 멀티미디어 카드 마이크로 타입(multimedia card micro type), 카드 타입의 메모리(예를 들어 SD 또는 XD 메모리 등), 램(Random Access Memory, RAM), SRAM(Static Random Access Memory), 롬(Read-Only Memory, ROM), EEPROM(Electrically Erasable Programmable Read-Only Memory), PROM(Programmable Read-Only Memory) 자기 메모리, 자기 디스크, 광디스크 중 적어도 하나의 타입의 저장매체를 포함할 수 있다. 아울러, 메모리는 일시적, 영구적 또는 반영구적으로 정보를 저장할 수 있으며, 내장형 또는 탈착형으로 제공될 수 있다.

[0071] 또한, 저장모듈(130)은 서버 네트워크를 통해 인공지능 기반 생성모델을 경량화하여 제공하기 위해 필요한 각종 정보를 저장하는 데이터베이스를 구축하거나, 별도의 외부 서버(클라우드 서버를 포함함)와 연동될 수도 있다.

[0072] 한편, 제어모듈(150)은 응용 프로그램과 관련된 동작 외에도, 적어도 하나의 프로세서를 기반으로 경량화 장치(100) 내 모든 구성들을 제어하여 입력 또는 출력되는 신호, 데이터, 정보 등을 처리하거나, 적어도 하나의 메모리에 저장된 명령어, 알고리즘, 응용 프로그램을 실행하여 각종 프로세스를 수행하며, 서버 네트워크를 통해 인공지능 기반 생성모델을 경량화하여 제공하기 위한 적절한 정보 또는 기능을 제공 또는 처리할 수 있다.

[0073] 구체적으로, 제어모듈(150)은 엡업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하고, 임의의 서버 네트워크를 찾은 후, 각각의 순방향 경로에서 할당된 스코어를 정렬하여 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트한다.

[0074] 이때, 제어모듈(150)은 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 다른 웨이트를 0으로 만들고, 역방향 경로에서는 상기 서버 네트워크의 로스를 산출하여 역전파를 활용하도록 한다.

[0075] 한편, 제어모듈(150)은 서버 네트워크의 로스를 산출할 시, 상기 서버 네트워크를 통해 생성된 이미지와 실제 이미지를 임베딩 스케이스로 보내 최대 평균 불일치(Maximum Mean Discrepancy, MMD) 스코어를 산출한다.

[0076] 이때, 제어모듈(150)은 두 개의 샘플 셋으로서 실제 샘플 $\{x_i\}_{i=1}^M$ 및 가짜 샘플 $\{f_i\}_{i=1}^M$ 에 대해 모든 차수의 모멘트를 매칭함으로써 최대 평균 불일치를 산출할 수 있다.

[0077] 구체적으로, 앞서 설명한 바와 같이, 최대 평균 불일치는 상기 <수학식 1>을 기초로 산출될 수 있다.

[0078] 도 7은 본 개시의 일 실시예에 따라 서버 네트워크를 통한 인공지능 기반 생성모델 경량화 방법을 나타내는 도면이다.

[0079] 도 7을 참조하면, 경량화 장치(100)는 엡업 알고리즘을 기반으로 텐스 네트워크에 대한 각각의 가중치에 랜덤 초기화된 스코어(s)를 할당하고(S210), 임의의 서버 네트워크를 찾는다(검색한다)(S220).

[0080] 그 다음으로, 경량화 장치(100)는 S220 단계에 의해 찾은 서버 네트워크에서 각각의 순방향 경로에서 할당된 스코어를 정렬하고(S230), 기 설정된 상위 k% 스코어를 가지는 가중치를 남기고 역전파를 활용하여 스코어를 업데이트 한다(S240).

[0081] 도 8은 본 개시에 따라 생성모델에서 SLTs가 존재하는지 여부를 확인하기 위해 진행한 실험 결과의 일 예시를 나타내는 도면으로서, 서버 네트워크 및 학습된 텐스 네트워크(Trained dense network)의 FID 스코어를 비교한 일 예시를 나타내는 도면이다(GFMN; LSUN-Bedroom).

[0082] 도 8을 참조하면, 가중치 업데이트 없이 랜덤 초기화된 신경망을 가지치기한다고 했을 때, 여기서, 가지치기된 서버 네트워크의 남아있는 가중치의 비율(%)인 다양한 k에 대해 FID를 시각화 한다.

[0083] 이때, 남아있는 가중치의 비율, 즉, 가중치를 얼마나 남길지에 대한 것들을 바꿔가면서 서버 네트워크를 구하고, 그 성능을 나타낸 것인데, FID 스코어는 생성모델의 성능을 나타내는 지표로, 낮을수록 좋은 지표이다.

[0084] 도 8에서 실선 라인은 학습된 모델의 성능을 나타내고, 점선 라인은 해당 알고리즘을 랜덤 네트워크에 적용하여 남은 웨이트의 양(k)에 따른 서버 네트워크의 성능 변화를 나타낸 것이다. 가 줄어들수록 좋은 성능을 보여주는 것을 알 수 있다. k가 높은 지점에서는 좋지 않은 성능을 보여주는데 이것은 k가 높을수록 랜덤 텐스 네트워크에 가깝기 때문에 자연스러운 현상이다. 반면에, k가 감소할수록 생성모델의 성능이 좋아지며 k가 10%에 도달했을 때 서버 네트워크의 성능이 학습된 텐스 네트워크와 겹치는 것을 확인할 수 있다.

- [0085] 결과적으로, 그 해당 지점을 SLTs로 볼 수 있다.
- [0086] 한편, 본 개시는 인공 신경망 방식으로 구현된 모델을 이용해서 소정의 목적 하에 예측(inference)을 수행하는 바, 이하에서는 인공 신경망에 대해 살펴보기로 한다.
- [0087] 본 명세서에서의 모델은 네트워크 함수, 인공신경망 및/또는 뉴럴 네트워크에 기반하여 동작하는 임의의 형태의 컴퓨터 프로그램을 의미할 수 있다. 본 명세서에 걸쳐, 모델, 신경망, 네트워크 함수, 뉴럴 네트워크(neural network)는 상호 교환 가능한 의미로 사용될 수 있다. 신경망은 하나 이상의 노드들이 하나 이상의 링크를 통해 상호 연결되어 신경망 내에서 입력 노드 및 출력 노드 관계를 형성한다. 신경망 내에서 노드들과 링크들의 개수 및 노드들과 링크들 사이의 연관관계, 링크들 각각에 부여된 가중치의 값에 따라, 신경망의 특성이 결정될 수 있다. 신경망은 하나 이상의 노드들의 집합으로 구성될 수 있다. 신경망을 구성하는 노드들의 부분 집합은 레이어(layer)를 구성할 수 있다.
- [0088] 딥 뉴럴 네트워크(DNN: deep neural network, 심층신경망)는 입력 레이어와 출력 레이어 외에 복수개의 히든 레이어를 포함하는 신경망을 의미할 수 있으며, 중간에 있는 히든 계층이 딥 뉴럴 네트워크에서는 1개 이상, 바람직하게는 2개 이상으로 구성된다.
- [0089] 이러한 딥 뉴럴 네트워크는 합성곱 신경망(CNN: convolutional neural network), 비전 트랜스포머(vision transformer), 리커런트 뉴럴 네트워크(RNN: recurrent neural network), LSTM(Long Short Term Memory) 네트워크, GPT(Generative Pre-trained Transformer), 오토 인코더(auto encoder), GAN(Generative Adversarial Networks), 제한 볼츠만 머신(RBM: restricted boltzmann machine), 심층 신뢰 네트워크(DBN: deep belief network), Q 네트워크, U 네트워크, 삼 네트워크, 적대적 생성 네트워크(GAN: Generative Adversarial Network), 트랜스포머(transformer) 등을 포함할 수 있다.
- [0090] 또는, 실시예에 따라 딥 뉴럴 네트워크는 전이학습(transfer learning) 방식으로 학습된 모델일 수 있다. 여기서, 전이학습은 대용량의 라벨링되어 있지 않은 학습용 데이터를 준지도학습 또는 자가학습 방식으로 사전 학습(pre-training)하여 제1 태스크를 갖는 사전 학습된(pre-trained) 모델(또는 베이스부)을 도 11 및 도 12에 도시되어 있는 기법(MLM 및 NSP) 등을 통해 얻고, 사전 학습된 모델을 제2 태스크에 적합하도록 fine-tuning하기 위해, 라벨링된 학습용 데이터를 지도학습 방식으로 학습시켜서 타겟으로 하는 모델을 구현하는 학습 방식을 나타낸다. 이러한 전이학습 방식으로 학습된 모델 중 하나로서, BERT(Bidirectional Encoder Representations from Transformers) 등이 있는데, 다만 이에 한정되는 것은 아니다.
- [0091] 전술한 딥 뉴럴 네트워크의 기재는 예시일 뿐이며 본 개시는 이에 제한되지 않는다. 여기서 전술한 합성곱 신경망의 경우, 이미지로부터 특징을 추출하는 특징 추출부(feature learning), 그리고 이렇게 추출된 특징을 이용해서 분류를 수행하는 분류부(classification)로 구성된다. 특징 추출부에는 이미지로부터 커널을 이용해서 특징이 추출되는 합성곱 계층, 활성화 함수 중 하나인 ReLU 계층 그리고 데이터의 차원을 줄이기 위한 풀링(Pooling) 계층이 포함될 수 있으며, 다만 이에 한정되는 것은 아니다. 아울러, 분류부에는 특징 추출부에서 추출된 특징을 일렬로 늘어뜨리는 flatten 계층, 그리고 실질적으로 분류가 수행되는 전결합층(fully connected layer) 및 softmax 함수가 포함될 수 있으며, 다만 이에 한정되는 것은 아니다.
- [0092] 뉴럴 네트워크는 지도학습(supervised learning), 비지도학습(unsupervised learning), 준지도학습(semi supervised learning), 자가학습(self-supervised learning) 또는 강화학습(reinforcement learning) 중 적어도 하나의 방식으로 학습될 수 있다. 뉴럴 네트워크의 학습은 뉴럴 네트워크가 특정한 동작을 수행하기 위한 지식을 뉴럴 네트워크에 적용하는 과정일 수 있다.
- [0093] 뉴럴 네트워크는 출력의 오류를 최소화하는 방향으로 학습될 수 있다. 뉴럴 네트워크의 학습에서 반복적으로 학습 데이터를 뉴럴 네트워크에 입력시키고 학습 데이터에 대한 뉴럴 네트워크의 출력과 타겟의 에러를 계산하고, 에러를 줄이기 위한 방향으로 뉴럴 네트워크의 에러를 뉴럴 네트워크의 출력 레이어에서부터 입력 레이어 방향으로 역전파(backpropagation)하여 뉴럴 네트워크의 각 노드의 가중치를 업데이트 하는 과정이다. 지도학습의 경우 각각의 학습 데이터에 정답이 라벨링 되어있는 데이터(labelled data)를 사용하며, 비지도학습의 경우는 각각의 학습 데이터에 정답이 라벨링되어 있지 않은 데이터(unlabeled data)를 사용할 수 있다. 업데이트 되는 각 노드의 연결 가중치는 학습률(learning rate)에 따라 변화량이 결정될 수 있다. 입력 데이터에 대한 뉴럴 네트워크의 계산과 에러의 역전파는 학습 사이클(epoch)을 구성할 수 있다. 학습률은 뉴럴 네트워크의 학습 사이클의 반복 횟수에 따라 상이하게 적용될 수 있다. 또한, 과적합(overfitting)을 막기 위해서 학습 데이터의 증가, 레귤러화(regularization), 노드 일부를 비활성화하는 드롭아웃(dropout), 배치 정규화 레이어(batch

normalization layer) 등의 방법이 적용될 수 있다.

- [0094] 한편, 일 실시예에서 개시되는 모델은 트랜스포머의 적어도 일부분을 차용할 수 있다. 트랜스포머는 임베딩된 데이터들을 인코딩하는 인코더 및 인코딩된 데이터들을 디코딩하는 디코더로 구성될 수 있다. 트랜스포머는 일련의 데이터들을 수신하여, 인코딩 및 디코딩 단계를 거쳐 상이한 타입의 일련의 데이터들을 출력하는 구조를 지닐 수 있다. 일 실시예에서, 일련의 데이터들은 트랜스포머가 연산가능한 형태로 가공될 수 있다. 일련의 데이터들을 트랜스포머가 연산가능한 형태로 가공하는 과정은 임베딩 과정을 포함할 수 있다. 데이터 토큰, 임베딩 벡터, 임베딩 토큰 등과 같은 표현들은, 트랜스포머가 처리할 수 있는 형태로 임베딩된 데이터들을 지칭하는 것일 수 있다.
- [0095] 트랜스포머가 일련의 데이터들을 인코딩 및 디코딩하기 위하여, 트랜스포머 내의 인코더 및 디코더들을 어텐션(attention) 알고리즘을 활용하여 처리할 수 있다. 어텐션 알고리즘이란 주어진 쿼리(Query)에 대해, 하나 이상의 키(Key)에 대한 유사도를 구하고, 이렇게 주어진 유사도를, 각각의 키(Key)와 대응하는 값(Value)에 반영한 후, 유사도가 반영된 값(Value)들을 가중합하여 어텐션 값을 계산하는 알고리즘을 의미할 수 있다.
- [0096] 쿼리, 키 및 값을 어떻게 설정하느냐에 따라, 다양한 종류의 어텐션 알고리즘이 분류될 수 있다. 예를 들어, 쿼리, 키 및 값을 모두 동일하게 설정하여 어텐션을 구하는 경우, 이는 셀프-어텐션 알고리즘을 의미할 수 있다. 입력된 일련의 데이터들을 병렬로 처리하기 위해, 임베딩 벡터를 차원을 축소하여, 각 분할된 임베딩 벡터에 대해 개별적인 어텐션 헤드를 구하여 어텐션을 구하는 경우, 이는 멀티-헤드(multi-head) 어텐션 알고리즘을 의미할 수 있다.
- [0097] 일 실시예에서, 트랜스포머는 복수개의 멀티-헤드 셀프 어텐션 알고리즘 또는 멀티-헤드 인코더-디코더 알고리즘을 수행하는 모듈들로 구성될 수 있다. 일 실시예에서, 트랜스포머는 임베딩, 정규화, 소프트맥스(softmax) 등 어텐션 알고리즘이 아닌 부가적인 구성요소들 또한 포함할 수 있다. 어텐션 알고리즘을 이용하여 트랜스포머를 구성하는 방법은 Vaswani et al., Attention Is All You Need, 2017 NIPS에 개시된 방법을 포함할 수 있으며, 이는 본 명세서에 참조로 통합된다.
- [0098] 트랜스포머는 임베딩된 자연어, 분할된 이미지 데이터, 오디오 파형 등 다양한 데이터 도메인에 적용하여, 일련의 입력 데이터를 일련의 출력 데이터로 변환할 수 있다. 다양한 데이터 도메인을 가진 데이터들을 트랜스포머에 입력가능한 일련의 데이터들로 변환하기 위해, 트랜스포머는 데이터들을 임베딩할 수 있다. 트랜스포머는 일련의 입력 데이터 사이의 상대적 위치관계 또는 위상관계를 표현하는 추가적인 데이터를 처리할 수 있다. 또는 일련의 입력 데이터에 입력 데이터들 사이의 상대적인 위치관계 또는 위상관계를 표현하는 벡터들이 추가적으로 반영되어 일련의 입력 데이터가 임베딩될 수 있다. 일 예에서, 일련의 입력 데이터 사이의 상대적 위치관계는, 자연어 문장 내에서의 어순, 각각의 분할된 이미지의 상대적 위치 관계, 분할된 오디오 파형의 시간 순서 등을 포함할 수 있으나, 이에 제한되지 않는다. 일련의 입력 데이터들 사이의 상대적인 위치관계 또는 위상관계를 표현하는 정보를 추가하는 과정은 위치 인코딩(positional encoding)으로 지칭될 수 있다.
- [0099] 상기 기술한 프로그램은, 상기 컴퓨터가 프로그램을 읽어 들여 프로그램으로 구현된 상기 방법들을 실행시키기 위하여, 상기 컴퓨터의 프로세서(CPU)가 상기 컴퓨터의 장치 인터페이스를 통해 읽힐 수 있는 C, C++, JAVA, 기 제어 등의 컴퓨터 언어로 코드화된 코드(Code)를 포함할 수 있다. 이러한 코드는 상기 방법들을 실행하는 필요한 기능들을 정의한 함수 등과 관련된 기능적인 코드(Functional Code)를 포함할 수 있고, 상기 기능들을 상기 컴퓨터의 프로세서가 소정의 절차대로 실행시키는데 필요한 실행 절차 관련 제어 코드를 포함할 수 있다. 또한, 이러한 코드는 상기 기능들을 상기 컴퓨터의 프로세서가 실행시키는데 필요한 추가 정보나 미디어가 상기 컴퓨터의 내부 또는 외부 메모리의 어느 위치(주소 번지)에서 참조되어야 하는지에 대한 메모리 참조관련 코드를 더 포함할 수 있다. 또한, 상기 컴퓨터의 프로세서가 상기 기능들을 실행시키기 위하여 원격(Remote)에 있는 어떠한 다른 컴퓨터나 서버 등과 통신이 필요한 경우, 코드는 상기 컴퓨터의 통신 모듈을 이용하여 원격에 있는 어떠한 다른 컴퓨터나 서버 등과 어떻게 통신해야 하는지, 통신 시 어떠한 정보나 미디어를 송수신해야 하는지 등에 대한 통신 관련 코드를 더 포함할 수 있다.
- [0100] 상기 저장되는 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, 상기 저장되는 매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광 데이터 저장장치 등이 있지만, 이에 제한되지 않는다. 즉, 상기 프로그램은 상기 컴퓨터가 접속할 수 있는 다양한 서버 상의 다양한 기록매체 또는 사용자의 상기 컴퓨터상의 다양한 기록매체에 저장될 수 있다. 또한, 상기 매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장될 수 있다.

[0101] 본 개시의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 개시가 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

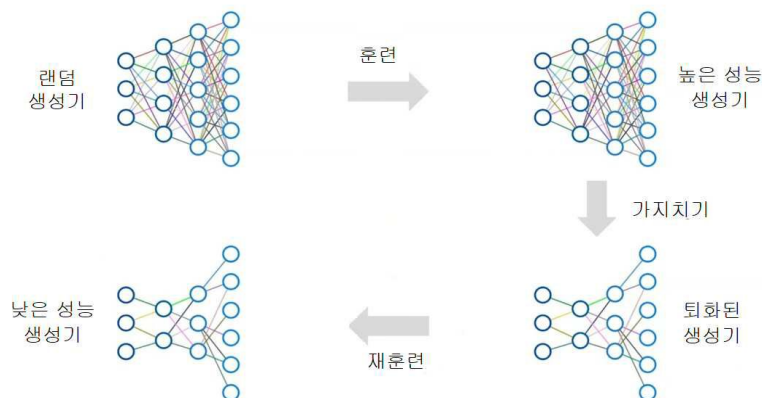
[0102] 이상, 첨부된 도면을 참조로 하여 본 개시의 실시예를 설명하였지만, 본 개시가 속하는 기술분야의 통상의 기술자는 본 개시가 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이지 않은 것으로 이해해야만 한다.

부호의 설명

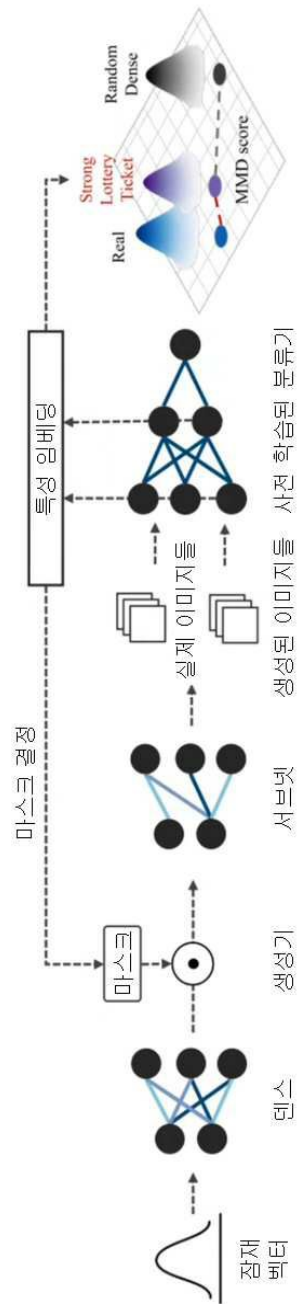
[0103]	100 : 경량화 장치	200 : 사용자 단말
	110 : 통신모듈	130 : 저장모듈
	150 : 제어모듈	

도면

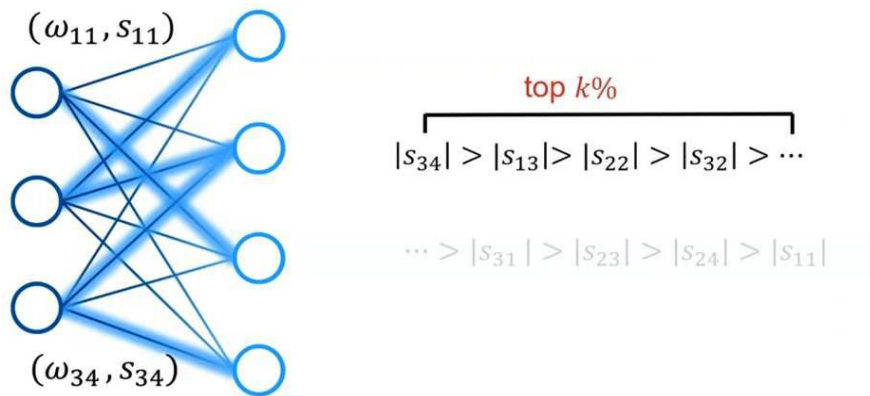
도면1



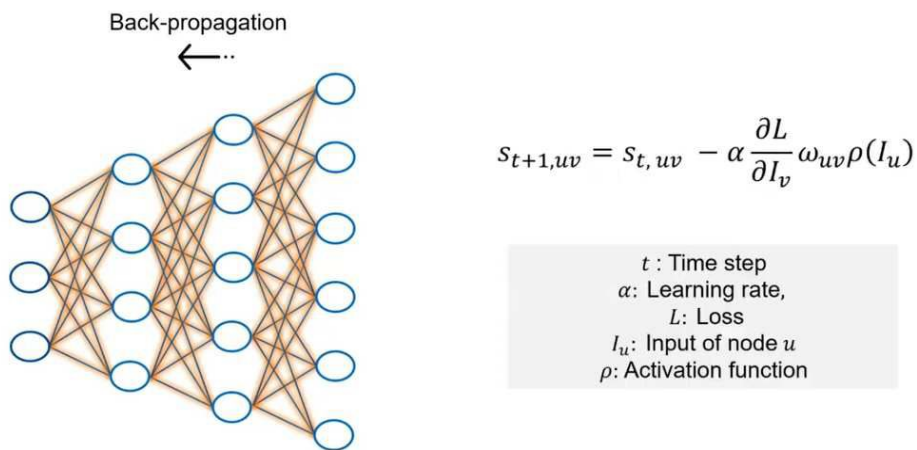
도면2



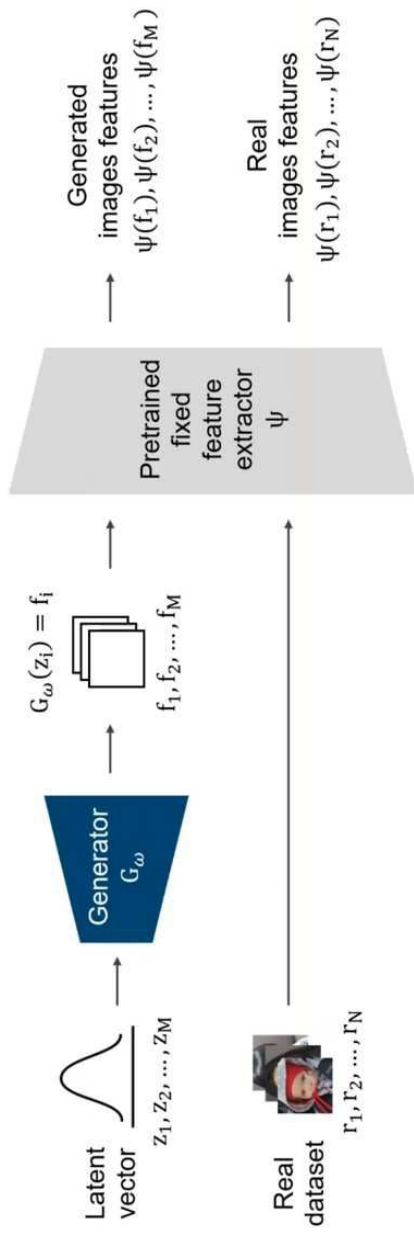
도면3



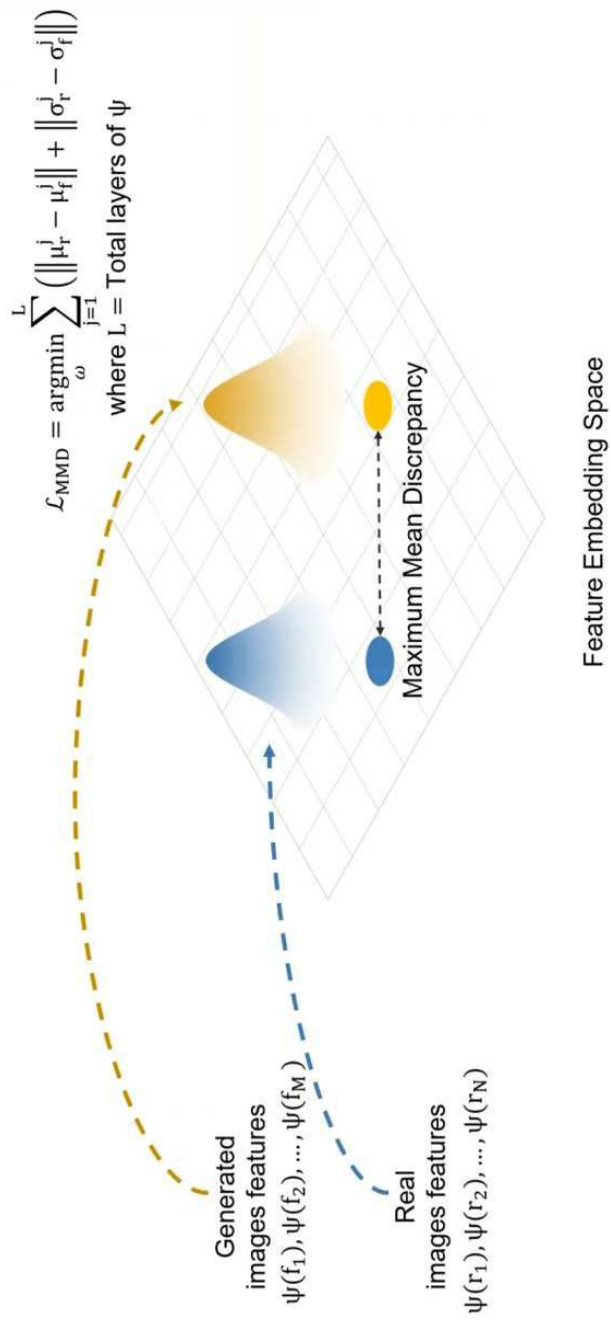
도면4



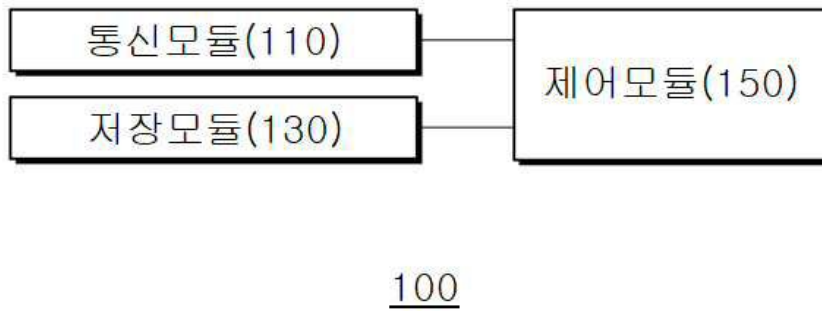
도면5



도면6



도면7



도면8

