



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0165601
(43) 공개일자 2024년11월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/098 (2023.01) G06N 3/0455 (2023.01)
G06N 3/0495 (2023.01)
(52) CPC특허분류
G06N 3/098 (2023.01)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0062927
(22) 출원일자 2023년05월16일
심사청구일자 2023년05월16일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
이훈
부산광역시 남구 유엔로 246 대연자이아파트 108동 2203호
이도윤
부산광역시 남구 이기대공원로26번길 45-3, 102-1611
(74) 대리인
이철희

전체 청구항 수 : 총 14 항

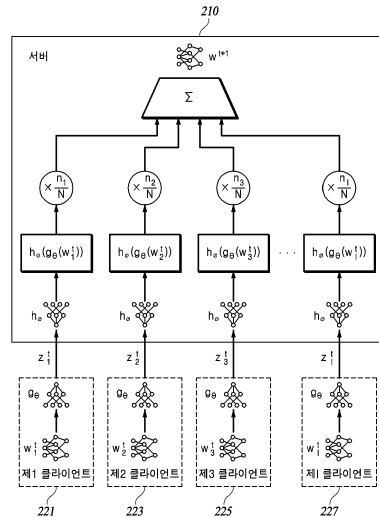
(54) 발명의 명칭 오토인코더 기반 통신효율적 연합학습을 위한 방법 및 장치

(57) 요약

오토인코더 기반 통신효율적 연합학습을 위한 방법 및 장치를 개시한다.

본 발명의 일 측면에 의하면, 오토인코더 기반 연합학습을 위한 컴퓨터 구현 방법에 있어서, 클라이언트들로부터 복수의 압축된 지역 모델을 수신하는 단계 - 상기 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축함 -; 디코더를 이용하여 상기 압축된 지역 모델들로부터 디코딩 결과들을 획득하는 단계; 상기 디코딩 결과들을 기반으로 전역 모델을 생성하는 단계; 및 상기 전역 모델을 상기 클라이언트들에게 전송하는 단계를 포함하는 방법 및 장치를 제공한다.

대표도 - 도2



(52) CPC특허분류

G06N 3/0495 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711134433
과제번호	2021-0-00467-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	6G 핵심기술개발
연구과제명	지능형 6G 무선 액세스 시스템
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

오토인코더 기반 연합학습을 위한 컴퓨터 구현 방법에 있어서,
클라이언트들로부터 복수의 압축된 지역 모델들을 수신하는 단계 - 상기 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축함 -;
디코더를 이용하여 상기 압축된 지역 모델들로부터 디코딩 결과들을 획득하는 단계;
상기 디코딩 결과들을 기반으로 전역 모델을 생성하는 단계; 및
상기 전역 모델을 상기 클라이언트들에게 전송하는 단계
를 포함하는 방법.

청구항 2

제1항에 있어서,
상기 전역 모델을 생성하는 단계는,
각 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 디코딩 결과를 가중하는 단계; 및
상기 가중된 디코딩 결과들을 평균하여 상기 전역 모델을 생성하는 단계
를 포함하는 방법.

청구항 3

제1항에 있어서,
상기 인코더 및 상기 디코더는,
서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 상기 인코더를 이용하여 압축하는 단계;
상기 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득하는 단계; 및
상기 트레이닝 지역 모델들과 상기 트레이닝 디코딩 결과들 간 차이값들의 평균을 기반으로 상기 인코더 및 상기 디코더를 업데이트하는 단계
를 포함하는 훈련 방법에 따라 훈련된 것인 방법.

청구항 4

제1항에 있어서,
상기 인코더 및 상기 디코더는,
서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 상기 인코더를 이용하여 압축하는 단계;
상기 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득하는 단계;
상기 디코딩 결과들을 기반으로 트레이닝 전역 모델을 생성하는 단계;
상기 트레이닝 지역 모델들을 기반으로 상기 트레이닝 전역 모델의 레이블을 생성하는 단계; 및
상기 트레이닝 전역 모델 및 상기 트레이닝 전역 모델의 레이블을 기반으로 상기 인코더 및 상기 디코더를 업데이트하는 단계
를 포함하는 훈련 방법에 따라 훈련된 것인 방법.

청구항 5

제4항에 있어서,

상기 트레이닝 전역 모델을 생성하는 단계는,

각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 디코딩 결과를 가중하는 단계; 및

상기 가중된 트레이닝 디코딩 결과들을 평균하여 상기 트레이닝 전역 모델을 생성하는 단계를 포함하는 방법.

청구항 6

제4항에 있어서,

상기 트레이닝 전역 모델의 레이블을 생성하는 단계는,

각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 지역 모델을 가중하는 단계; 및

상기 가중된 트레이닝 지역 모델들을 평균하여 상기 트레이닝 전역 모델의 레이블을 생성하는 단계를 포함하는 방법.

청구항 7

제4항에 있어서,

상기 인코더 및 상기 디코더를 업데이트하는 단계는,

상기 트레이닝 전역 모델 및 상기 트레이닝 전역 모델의 레이블 간 차이가 감소하도록 상기 인코더 및 상기 디코더를 업데이트하는 단계

를 포함하는 방법.

청구항 8

오토인코더 기반 연합학습을 위한 장치에 있어서,

명령어들을 저장하는 메모리; 및

적어도 하나의 프로세서를 포함하되,

상기 적어도 하나의 프로세서는 상기 명령어들을 실행함으로써,

클라이언트들로부터 복수의 압축된 지역 모델들을 수신하고 - 상기 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축함 -;

디코더를 이용하여 상기 압축된 지역 모델들로부터 디코딩 결과들을 획득하고,

상기 디코딩 결과들을 기반으로 전역 모델을 생성하고,

상기 전역 모델을 상기 클라이언트들에게 전송하는 장치.

청구항 9

제8항에 있어서,

상기 적어도 하나의 프로세서는,

각 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 디코딩 결과를 가중하고,

상기 가중된 디코딩 결과들을 평균하여 상기 전역 모델을 생성하는 장치.

청구항 10

제8항에 있어서,
 상기 인코더 및 상기 디코더는,
 서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 상기 인코더를 이용하여 압축하는 단계;
 상기 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득하는 단계; 및
 상기 트레이닝 지역 모델들과 상기 트레이닝 디코딩 결과들 간 차이값들의 평균을 기반으로 상기 인코더 및 상기 디코더를 업데이트하는 단계
 를 포함하는 훈련 방법에 따라 훈련된 것인 장치.

청구항 11

제8항에 있어서,
 상기 인코더 및 상기 디코더는,
 서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 상기 인코더를 이용하여 압축하는 단계;
 상기 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득하는 단계;
 상기 디코딩 결과들을 기반으로 트레이닝 전역 모델을 생성하는 단계;
 상기 트레이닝 지역 모델들을 기반으로 상기 트레이닝 전역 모델의 레이블을 생성하는 단계; 및
 상기 트레이닝 전역 모델 및 상기 트레이닝 전역 모델의 레이블을 기반으로 상기 인코더 및 상기 디코더를 업데이트하는 단계
 를 포함하는 훈련 방법에 따라 훈련된 것인 장치.

청구항 12

제11항에 있어서,
 상기 트레이닝 전역 모델을 생성하는 단계는,
 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 디코딩 결과를 가중하는 단계; 및
 상기 가중된 트레이닝 디코딩 결과들을 평균하여 상기 트레이닝 전역 모델을 생성하는 단계
 를 포함하는 장치.

청구항 13

제11항에 있어서,
 상기 트레이닝 전역 모델의 레이블을 생성하는 단계는,
 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 지역 모델을 가중하는 단계; 및
 상기 가중된 트레이닝 지역 모델들을 평균하여 상기 트레이닝 전역 모델의 레이블을 생성하는 단계
 를 포함하는 장치.

청구항 14

제11항에 있어서,
 상기 인코더 및 상기 디코더를 업데이트하는 단계는,
 상기 트레이닝 전역 모델 및 상기 트레이닝 전역 모델의 레이블 간 차이가 감소하도록 상기 인코더 및 상기 디코더를 업데이트하는 단계

를 포함하는 장치.

발명의 설명

기술 분야

[0001] 본 발명의 실시예들은 오토인코더 기반 통신효율적 연합학습을 위한 방법 및 장치에 관한 것이다.

배경 기술

[0002] 이하에 기술되는 내용은 단순히 본 실시예와 관련되는 배경 정보만을 제공할 뿐 종래기술을 구성하는 것이 아니다.

[0003] 최근 분산된 클라이언트들의 인공지능 모델을 원격으로 최적화하는 연합학습 시스템에 대한 연구가 활발히 진행되고 있다. 연합학습 시스템에 따르면, 서버는 클라이언트의 훈련 데이터가 아니라, 클라이언트들의 지역(local) 모델들을 취합한다. 훈련 데이터를 공유하지 않으므로 개인 정보 유출을 방지하고, 통신 효율적인 분산 훈련이 가능하다.

[0004] 하지만, 복잡한 문제를 해결하기 위해 모델의 크기, 즉 모델의 파라미터 수가 증가함에 따라, 서버와 클라이언트들 간 모델 교환을 위해 요구되는 통신 자원도 증가하였다.

[0005] 연합학습에 요구되는 통신 자원량을 줄이기 위해 통신 효율적 연합학습 시스템들이 연구되고 있다. 예를 들면, 연합학습 시스템은 인공지능 모델의 교환 전에 인공지능 모델의 파라미터들을 저차원 벡터로 선형 압축함으로써, 통신 자원량을 줄일 수 있게 되었다. 나아가, 연합학습 시스템은 고차원 인공지능 모델의 비선형적인 은닉 특성들을 압축하기 위해 오토인코더 신경망 모델을 채용하였다. 이를 통해, 연합학습 시스템은 고차원 인공지능 모델의 파라미터들의 은닉 특성을 효율적으로 추출하고, 복원할 수 있게 되었다.

[0006] 다만, 종래의 오토인코더 기반 연합학습 시스템에 따르면, 서버 및 클라이언트는 서로 다른 오토인코더들을 이용한다. 복수의 오토인코더들이 훈련된 후에, 클라이언트들에는 인코더들이 각각 설치되고, 서버에는 인코더들에 대응되는 복수의 디코더들이 설치된다. 여기서, 오토인코더들의 개수는 클라이언트들의 고정된 개수에 따라 결정된다. 다시 말하면, 훈련된 오토인코더들의 개수와 클라이언트들의 수가 달라지면, 연합학습 시스템의 훈련 성능이 현저하게 악화된다.

[0007] 나아가, 서로 다른 오토인코더들은 서로 다른 데이터 분포를 갖는 지역 모델들을 압축 및 압축해제하도록 훈련되기 때문에, 각 인코더는 특정 데이터 분포에 과적합될 가능성이 높다. 예를 들면, 어느 지역 모델의 파라미터들의 확률분포가 iid(non-independent and identically distributed) 특성을 갖는다면, 대응되는 오토인코더는 iid 특성에 따른 추론 데이터에 대해 좋은 복원 성능을 나타낼 수 있는 반면, non-iid 특성에 따른 추론 데이터에서는 성능이 악화될 수 있다.

[0008] 또한, 종래의 오토인코더 기반 연합학습 시스템에서 오토인코더는 인공지능 모델을 압축하고, 압축된 모델로부터 인공지능 모델을 복원하도록 훈련된다. 이때, 오토인코더는 인공지능 모델을 최대한 동일하게 복원하도록 훈련된다. 하지만, 연합학습 시스템에서 궁극적인 목표는 지역 모델들로부터 전역 모델을 생성하는 것이다. 즉, 오토인코더의 복원 성능만을 향상시키는 것은 연합학습되는 전역 모델의 성능을 향상시키는 데 한계가 있다.

[0009] 따라서, 오토인코더 기반 연합학습 시스템에서 오토인코더들의 개수와 과적합에 따른 문제를 해결하기 위한 방안이 필요하다. 나아가, 오토인코더의 훈련 방법에 대한 고민도 필요하다.

발명의 내용

해결하려는 과제

[0010] 본 발명의 실시예들은, 서버와 클라이언트들이 하나의 오토인코더의 파라미터들을 공유함으로써 클라이언트들의 수의 변화에 대응가능하며, 오토인코더가 지역 모델의 특정 데이터 분포에 과적합되는 것을 방지하기 위한 연합학습 방법 및 장치를 제공하는 데 주된 목적이 있다.

[0011] 본 발명의 다른 실시예들은, 오토인코더에게 지역 모델의 복원과 함께 연합학습의 내재적 의미를 학습시켜 오토인코더의 일반화 능력을 개선함으로써, 전역모델의 최적화 속도 및 정확성을 개선하기 위한 연합학습 방법 및 장치를 제공하는 데 일 목적이 있다.

[0012] 본 발명이 해결하고자 하는 과제들은 이상에서 언급한 과제들로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0013] 본 발명의 일 측면에 의하면, 오토인코더 기반 연합학습을 위한 컴퓨터 구현 방법에 있어서, 클라이언트들로부터 복수의 압축된 지역 모델을 수신하는 단계 - 상기 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축함 -; 디코더를 이용하여 상기 압축된 지역 모델들로부터 디코딩 결과들을 획득하는 단계; 상기 디코딩 결과들을 기반으로 전역 모델을 생성하는 단계; 및 상기 전역 모델을 상기 클라이언트들에게 전송하는 단계를 포함하는 방법을 제공한다.

[0014] 본 실시예의 다른 측면에 의하면, 오토인코더 기반 연합학습을 위한 장치에 있어서, 명령어들을 저장하는 메모리; 및 적어도 하나의 프로세서를 포함하되, 상기 적어도 하나의 프로세서는 상기 명령어들을 실행함으로써, 클라이언트들로부터 복수의 압축된 지역 모델들을 수신하고 - 상기 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축함 -; 디코더를 이용하여 상기 압축된 지역 모델들로부터 디코딩 결과들을 획득하고, 상기 디코딩 결과들을 기반으로 전역 모델을 생성하고, 상기 전역 모델을 상기 클라이언트들에게 전송하는 장치를 제공한다.

발명의 효과

[0015] 이상에서 설명한 바와 같이 본 발명의 일 실시예에 의하면, 서버와 클라이언트들이 하나의 오토인코더의 파라미터들을 공유함으로써 클라이언트들의 수의 변화에 대응가능하며, 오토인코더가 지역 모델의 특정 데이터 분포에 과적합되는 것을 방지할 수 있다.

[0016] 본 발명의 다른 실시예에 의하면, 오토인코더에게 지역 모델의 복원과 함께 연합학습의 내재적 의미를 학습시켜 오토인코더의 일반화 능력을 개선함으로써, 전역모델의 최적화 속도 및 정확성을 개선할 수 있다.

[0017] 본 발명의 다른 실시예에 의하면, 전역 모델의 최적화 속도를 높임으로써, 서버와 클라이언트들 간 모델의 교환 횟수 및 통신 자원요구량을 줄일 수 있다.

[0018] 본 개시의 효과들은 이상에서 언급한 효과들로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0019] 도 1은 연합학습 시스템의 구성도이다.

도 2는 본 발명의 일 실시예에 따른 연합학습 시스템의 구성도이다.

도 3은 본 발명의 일 실시예에 따른 오토인코더의 전역 모델 기반 훈련을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 연합학습의 순서도이다.

도 5는 본 발명의 일 실시예에 따른 오토인코더의 훈련 방법의 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 이하, 본 개시의 일부 실시예들을 예시적인 도면을 이용해 상세하게 설명한다. 각 도면의 구성 요소들에 참조 부호를 부가함에 있어서, 동일한 구성 요소들에 대해서는 비록 다른 도면 상에 표시되더라도 가능한 한 동일한 부호를 가지도록 하고 있음에 유의해야 한다. 또한, 본 개시를 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 개시의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다.

[0021] 본 개시에 따른 실시예의 구성요소를 설명하는 데 있어서, 제1, 제2, i), ii), a), b) 등의 부호를 사용할 수 있다. 이러한 부호는 그 구성요소를 다른 구성 요소와 구별하기 위한 것일 뿐, 그 부호에 의해 해당 구성요소의 본질 또는 차례나 순서 등이 한정되지 않는다. 명세서에서 어떤 부분이 어떤 구성요소를 '포함' 또는 '구비'한다고 할 때, 이는 명시적으로 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0022] 본 발명에 따른 장치 또는 방법의 각 구성요소는 하드웨어 또는 소프트웨어로 구현되거나, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다. 또한, 각 구성요소의 기능이 소프트웨어로 구현되고 마이크로프로세서가 각

구성요소에 대응하는 소프트웨어의 기능을 실행하도록 구현될 수도 있다.

[0023] 도 1은 연합학습 시스템의 구성도이다.

[0024] 도 1을 참조하면, 연합학습 시스템은 서버(110) 및 복수의 클라이언트들(121, 123, 125, 127)을 포함한다. 서버(110) 및 클라이언트들(121, 123, 125, 127)은 각각 신경망을 실행할 수 있는 컴퓨팅 장치로서, 데스크탑 PC, 노트북 PC, 태블릿 PC 등과 같이, 다양한 전자 장치로 구현될 수 있다.

[0025] 연합학습 시스템에서는 서버(110)와 클라이언트들(121, 123, 125, 127)이 협력적으로 인공지능 모델, 즉 신경망을 훈련한다. 기본적으로, 클라이언트들(121, 123, 125, 127)에서 훈련된 지역 모델들은 서버(110)로 전송되며, 서버(110)에서 지역 모델들로부터 전역 모델이 생성된다.

[0026] 구체적으로, 클라이언트들(121, 123, 125, 127)은 각각 훈련 데이터를 수집하고, 훈련 데이터를 기반으로 지역 모델을 훈련한다. 여기서, 지역 모델들은 서로 동일한 아키텍처를 가질 수 있다. 반면, 클라이언트들(121, 123, 125, 127)에서는 서로 다른 훈련 데이터가 수집되며, 다른 훈련 데이터를 기반으로 훈련된 지역 모델들은 서로 다른 파라미터 값들을 갖는다. 예를 들면, t번째 반복(iteration)에서, 제1 클라이언트(121)는 제1 지역 모델 w_1^t 을 생성하고, 제2 클라이언트(123)는 제2 지역 모델 w_2^t 을 생성한다. 여기서, 제1 지역 모델과 제2 지역 모델은 서로 다른 훈련 데이터를 훈련하기 때문에 서로 다른 지역 파라미터들을 갖는다. t번째 반복은 서버(110)와 클라이언트들(121, 123, 125, 127) 간 모델 교환 횟수를 나타낸다.

[0027] 한편, 지역 모델은 지도 학습, 비지도 학습 또는 강화학습 등 다양한 학습방법에 따라 훈련될 수 있다. 예를 들면, 지역 모델은 MSE(Mean Square Error) 기반 손실함수와 SGD(Stochastic Gradient Descent) 기반 업데이트를 이용하여 훈련될 수 있다.

[0028] 이후, 클라이언트들(121, 123, 125, 127)은 훈련한 지역 모델들을 서버(110)로 전송한다. 클라이언트들(121, 123, 125, 127)은 지역 모델들을 상향링크 통신채널을 통해 전송할 수 있다.

[0029] 서버(110)는 지역 모델들을 수신하고, 지역 모델들로부터 전역 모델을 생성한다. 구체적으로, 서버(110)는 지역 모델들의 평균 또는 가중평균을 전역 모델로서 생성한다. t번째 반복에 따른 지역 모델들은 t+1번째 반복에 대한 전역 모델 w^{t+1} 로 취합된다.

[0030] 특히, 서버(110)는 각 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 이용하여 지역 모델들의 가중평균을 계산할 수 있다. 훈련한 샘플수가 많은 지역 모델은 전역 모델 생성에 많은 영향을 미친다. 구체적으로, 서버(110)는 수학식 1을 기반으로 지역 모델들의 가중평균을 계산할 수 있다.

수학식 1

$$w^{t+1} = \sum_{i=1}^I \frac{n_i}{N} w_i^t$$

[0031]

[0032] 수학식 1에서, t는 반복 횟수, w^{t+1} 은 전역 모델, w_i^t 은 i번째 지역 모델, i는 지역 모델의 인덱스, I는 클라이언트들의 수, N은 지역 모델들에 훈련된 데이터 샘플들의 총 개수, n_i 는 i번째 지역 모델에 훈련된 데이터 샘플들의 개수를 나타낸다.

[0033] 서버(110)는 클라이언트들(121, 123, 125, 127)로 하여금 전역 모델을 다시 지역 모델로 이용하도록 전역 모델을 클라이언트들(121, 123, 125, 127)에게 배포한다. 서버(110)는 하향링크 통신채널을 통해 전역 모델을 전송할 수 있다. 클라이언트들(121, 123, 125, 127)은 전역 모델을 초기 모델로 설정하고, 서로 다른 훈련 데이터를 다시 수집하고, 수집된 훈련 데이터를 기반으로 전역 모델을 각자 업데이트함으로써, 지역 모델을 생성한다.

[0034] 전술한 과정을 반복함으로써, 연합학습 시스템은 클라이언트들(121, 123, 125, 127)의 개인정보와 같은 프라이버시를 침해하지 않은 채 인공지능 모델을 최적화할 수 있다.

[0035] 다만, 지역 모델 및 전역 모델의 파라미터 수에 따라 서버(110)와 클라이언트들(121, 123, 125, 127) 간 많은

통신 비용이 발생할 수 있다. 또한, 통신 지연으로 인해 연합학습에 소요되는 훈련 시간이 기하급수적으로 증가할 수 있다.

[0036] 이러한 문제점을 해결하기 위해, 연합학습 시스템은 오토인코더를 이용함으로써, 요구되는 통신 자원 및 훈련 시간을 줄일 수 있다.

[0037] 도 2는 본 발명의 일 실시예에 따른 연합학습 시스템의 구성도이다.

[0038] 도 2를 참조하면, 연합학습 시스템은 서버(210) 및 복수의 클라이언트들(221, 223, 225, 227)을 포함한다. 연합학습 시스템은 연합학습을 위한 통신 비용 및 학습 시간을 줄이기 위해 오토인코더를 이용한다.

[0039] 종래의 연합학습 시스템에 따르면, 클라이언트들에는 서로 다른 인코더들이 설치되고, 서버에는 클라이언트들의 인코더들에 대응되는 디코더들이 설치된다. 여기서, 인코더는 클라이언트의 지역 모델을 압축하고, 디코더는 압축된 지역 모델을 압축해제하도록 훈련되는 신경망 모델이다. 구체적으로, 인코더는 인공지능 모델을 높은 차원에서 낮은 차원으로 압축하도록 훈련되고, 디코더는 압축된 모델을 낮은 차원에서 높은 차원으로 디코딩하도록 훈련된다.

[0040] 이때, 클라이언트들의 개수만큼 오토인코더들이 미리 최적화되며, 클라이언트들의 개수가 증가하면 최적화되지 않은 오토인코더들이 추가된다. 이에 따라, 클라이언트들의 지역 모델들의 압축 및 복원이 제대로 수행되지 않았다.

[0041] 나아가, 각 오토인코더는 유사한 데이터 분포를 갖는 모델들을 압축 및 압축해제하도록 훈련된다. 즉, 각 오토인코더는 소정의 데이터 분포에 과적합될 수 있다. 각 오토인코더가 훈련한 데이터 분포와 다른 데이터 분포를 갖는 모델을 압축 및 압축해제 시, 성능 저하가 발생한다. 이 경우, 전역 모델 또한 최적화되기 어렵다.

[0042] 이와 달리, 본 발명의 일 실시예에 따른 연합학습 시스템은 하나의 훈련된 오토인코더를 이용한다. 구체적으로, 클라이언트들(221, 223, 225, 227)에는 훈련된 오토인코더로부터 인코더들(g_θ)이 복사 및 설치된다. 서버(210)에는 훈련된 오토인코더의 디코더(h_ϕ)가 설치된다. 도 2에서, 서버(210)는 인코더들에 대응되도록 복수의 디코더들을 포함하는 것으로 표현되지만, 서버(210)에는 하나의 디코더가 설치될 수 있다. 이처럼, 연합학습 시스템은 동일한 파라미터 값을 갖는 인코더들을 이용하여 지역 모델들 w_i^t 을 압축하며, 클라이언트의 수가 증가하더라도 인코더를 복사하여 추가된 클라이언트에 복사된 인코더를 포함시킴으로써, 클라이언트의 수에 영향을 받지 않을 수 있다.

[0043] 나아가, 연합학습 시스템에서 클라이언트들(221, 223, 225, 227)에 동일한 오토인코더가 적용되므로, 오토인코더가 특정 데이터 분포에 과적합되는 것을 방지할 수 있다. 이를 위해, 오토인코더는 클라이언트들(221, 223, 225, 227)의 지역 모델들을 모두 훈련한다. 지역 모델들은 다양한 데이터 분포들을 가지므로, 오토인코더는 지역 모델들의 다양한 데이터 분포를 훈련하여, 일반화 능력을 개선할 수 있다.

[0044] 이러한 오토인코더는 지역 모델 기반 손실함수 또는 전역 모델 기반 손실함수에 따라 훈련될 수 있다.

[0045] 지역 모델 기반 손실함수는 오토인코더가 지역 모델들을 압축하고, 압축된 지역 모델들을 복원하도록 훈련시키기 위한 함수이다. 오토인코더의 출력에 대한 레이블은 오토인코더의 입력이다. 지역 모델 기반 손실함수는 수학식 2과 같이 표현될 수 있다.

수학식 2

$$L_{AB} = \frac{1}{TK} \sum_{t=1}^T \sum_{i=1}^I \|w_i^t - h_\phi(g_\theta(w_i^t))\|^2$$

[0046]

[0047] 수학식 2에서, L_{AB} 는 손실함수, T 는 클라이언트-서버 간 모델의 총 교환 횟수, t 는 모델 교환의 반복 차수, i 는 클라이언트의 인덱스, I 는 클라이언트들의 전체 수, w_i^t 는 제 i 트레이닝 지역 모델, $h_\phi(g_\theta(w_i^t))$ 는 제 i 트

레이닝 지역 모델의 디코딩 결과를 나타낸다.

- [0048] 오토인코더는 입력과 동일한 출력을 생성하도록 훈련된다. 다만, 오토인코더는 다양한 데이터 분포를 훈련하기 위해, 트레이닝 지역 모델들과 트레이닝 디코딩 결과들 간 차이값들의 평균을 기반으로 훈련된다. 이 과정에서, 오토인코더는 다양한 데이터 분포에 따른 트레이닝 지역 모델들에 훈련될 수 있다.
- [0049] 한편, 전역 모델 기반 손실함수는 오토인코더에 지역 모델의 복원과 함께 연합학습의 내재적 의미를 학습시키기 위한 함수이다. 전역 모델 기반 손실함수는 디코딩 결과들로부터 도출되는 전역 모델과 지역 모델들로부터 도출되는 전역 모델 간 차이에 의해 정의된다. 즉, 오토인코더는 지역 모델과 디코딩 결과가 아니라, 지역 모델들에 따른 전역 모델과 디코딩 결과들에 따른 전역 모델들을 기반으로 훈련된다. 따라서, 서버(210) 내 디코더의 디코딩 결과는 대응되는 지역 모델과는 다르다. 이에 대해서는, 도 3에서 자세히 설명한다.
- [0050] 이하에서는, 연합학습 과정에 대해 설명한다.
- [0051] 먼저, 클라이언트들(221, 223, 225, 227)은 서버(210) 또는 외부 장치로부터 초기 지역 모델들을 수신하고, 훈련 데이터를 수집하고, 훈련 데이터를 기반으로 초기 지역 모델들을 업데이트함으로써, 지역 모델들을 생성한다.
- [0052] 클라이언트들(221, 223, 225, 227)은 각각 인코더를 이용하여 지역 모델을 압축하고, 압축된 지역 모델 z_i^t ($i=1, 2, 3, \dots, I$)을 서버(210)로 전송한다. 여기서, 각 압축된 지역 모델 z_i^t 은 $g_\theta(w_i^t)$ 와 같고, 통신 신호 형태로 전송될 수 있다.
- [0053] 서버(210)는 클라이언트들(221, 223, 225, 227)로부터 압축된 지역 모델들을 수신하고, 디코더를 이용하여 압축된 지역 모델들로부터 디코딩 결과들을 획득한다. 예를 들면, 서버(210)는 제1 압축된 지역 모델 z_1^t 을 디코딩하여 제1 디코딩 결과 $h_\phi(g_\theta(w_1^t))$ 를 획득하고, 제I 압축된 지역 모델 z_I^t 를 디코딩하여 제I 디코딩 결과 $h_\phi(g_\theta(w_I^t))$ 를 획득한다.
- [0054] 여기서, 클라이언트들(221, 223, 225, 227)의 인코더들과 서버(210)의 디코더가 지역 모델 기반 손실함수에 따라 훈련된 경우, 디코딩 결과는 압축된 지역 모델의 복원 결과이다. 반면, 인코더와 디코더가 전역 모델 기반 손실함수에 따라 훈련된 경우, 디코더의 디코딩 결과는 인코더에 입력되는 지역 모델과 다르다.
- [0055] 서버(210)는 디코딩 결과들을 기반으로 전역 모델 w^{t+1} 을 생성한다.
- [0056] 일 예로서, 서버(210)는 디코딩 결과들의 평균을 전역 모델로 생성할 수 있다.
- [0057] 다른 예로서, 서버(210)는 디코딩 결과들의 가중평균을 전역 모델로 생성할 수 있다. 구체적으로, 서버(210)는 각 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 디코딩 결과를 가중하고, 가중된 디코딩 결과들을 평균하여 전역 모델을 생성할 수 있다. 다시 말하면, 서버(210)는 수학적 식 3을 이용하여 디코딩 결과들의 가중평균을 계산할 수 있다.

수학적 식 3

$$w^{t+1} = \sum_{i=1}^I \frac{n_i}{N} h_\phi(g_\theta(w_i^t))$$

- [0058]
- [0059] 수학적 식 3에서, t 는 반복 횟수, w^{t+1} 은 전역 모델, w_i^t 은 i 번째 지역 모델, i 는 지역 모델의 인덱스, I 는 클라이언트들의 수, g_θ 는 인코더 함수, h_ϕ 는 디코더 함수, N 은 지역 모델들에 훈련된 데이터 샘플들의 총 개수, n_i 는 i 번째 지역 모델에 훈련된 데이터 샘플들의 개수를 나타낸다.
- [0060] 이후, 서버(210)는 생성된 전역 모델을 클라이언트들(221, 223, 225, 227)에게 하향링크 채널을 통해 전송한다.

클라이언트들(221, 223, 225, 227)은 전역 모델을 초기 지역 모델로 설정 및 업데이트함으로써, 다시 지역 모델을 생성한다.

[0061] 이처럼, 본 발명의 일 실시예에 따른 연합학습 시스템은 클라이언트들(221, 223, 225, 227)에 동일한 인코더들을 배포함으로써, 클라이언트 수가 훈련된 오토인코더 수보다 많아지더라도 전역 모델을 효율적으로 최적화할 수 있다. 즉, 수 많은 클라이언트들이 존재하는 환경에서, 전역 모델의 수렴 속도 및 정확도 성능이 모두 개선될 수 있다.

[0062] 도 3은 본 발명의 일 실시예에 따른 오토인코더의 전역 모델 기반 훈련을 설명하기 위한 도면이다.

[0063] 도 3을 참조하면, 오토인코더에 포함되는 인코더(310) 및 디코더(320)가 도시된다.

[0064] 오토인코더는 인코더들이 클라이언트들에 배포되고, 서버에 디코더가 포함된 형태로 훈련될 수 있으나, 별도의 훈련 장치 내에서 훈련될 수도 있다. 훈련 장치는 오토인코더의 훈련을 위한 컴퓨팅 장치이다. 훈련 장치는 클라이언트들의 트레이닝 지역 모델들을 수집하고, 트레이닝 지역 모델들을 이용하여 오토인코더를 훈련하고, 훈련된 오토인코더의 인코더 및 디코더를 각각 클라이언트와 서버에 전송할 수 있다.

[0065] 훈련 장치는 오토인코더가 연합학습의 내재적 의미를 훈련할 수 있도록 전역 모델 기반 손실함수를 이용하여 오토인코더를 훈련할 수 있다.

[0066] 구체적으로, 훈련 장치는 서로 다른 데이터 분포를 훈련한 트레이닝 지역 모델들 w_i^t 을 획득한다. 훈련 장치는 인코더(310)를 이용하여 트레이닝 지역 모델들을 압축하고, 디코더를 이용하여 압축된 트레이닝 지역 모델들 $g_\theta(w_i^t)$ 로부터 트레이닝 디코딩 결과들 $h_\phi(g_\theta(w_i^t))$ 을 획득한다. 인코더(310)는 트레이닝 지역 모델들을 압축하고, 디코더(320)는 압축된 트레이닝 지역 모델들을 디코딩하여 트레이닝 디코딩 결과들을 출력한다.

[0067] 이때, 트레이닝 지역 모델들은 오토인코더에 동시에 또는 순차적으로 입력될 수 있다.

[0068] 이후, 훈련 장치는 트레이닝 디코딩 결과들을 기반으로 트레이닝 전역 모델 w^{t+1} 을 생성한다. 훈련 장치는 트레이닝 디코딩 결과들의 평균 또는 가중평균을 트레이닝 전역 모델로 생성할 수 있다.

[0069] 특히, 훈련 장치는 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치 $\frac{n_i}{N}$ 로 각 트레이닝 디코딩 결과를 가중하고, 가중된 트레이닝 디코딩 결과들을 평균하여 트레이닝 전역 모델 $\hat{w}^{t+1}$ 을 생성한다. 자세하게는, 트레이닝 전역 모델은 수학적 식 3에 따라 생성될 수 있다.

[0070] 또한, 훈련 장치는 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 지역 모델을 가중하고, 가중된 트레이닝 지역 모델들을 평균하여 트레이닝 전역 모델의 레이블 w^t 을 생성한다. 자세하게는, 트레이닝 전역 모델의 레이블은 수학적 식 1에 따라 생성될 수 있다.

[0071] 훈련 장치는 트레이닝 전역 모델과 트레이닝 전역 모델의 레이블을 기반으로 인코더(310) 및 디코더(320)를 업데이트한다. 구체적으로, 훈련 장치는 트레이닝 전역 모델 및 트레이닝 전역 모델의 레이블 간 차이로 정의되는 전역 모델 기반 손실함수를 계산하고, 전역 모델 기반 손실함수가 감소하도록 인코더 및 디코더를 업데이트할 수 있다. 여기서, 전역 모델 기반 손실함수는 수학적 식 4와 같이 표현될 수 있다.

수학적 식 4

$$L_{FedAvgAE}(\theta, \phi) = \frac{1}{T} \sum_{t=1}^T \|w^{t+1} - \hat{w}^{t+1}\|^2 = \frac{1}{T} \sum_{t=1}^T \left\| \sum_{i=1}^I \frac{n_i}{N} w_i^t - \sum_{i=1}^I \frac{n_i}{N} h_\phi(g_\theta(w_i^t)) \right\|^2$$

[0072]

[0073] 수학적 식 4에서, $L_{FedAvgAE}$ 는 전역 모델 기반 손실함수, T 는 클라이언트-서버 간 모델의 총 교환 횟수, t 는 모델 교환의 반복 차수, i 는 클라이언트의 인덱스, I 는 클라이언트들의 전체 수, $\hat{w}^{t+1}$ 와

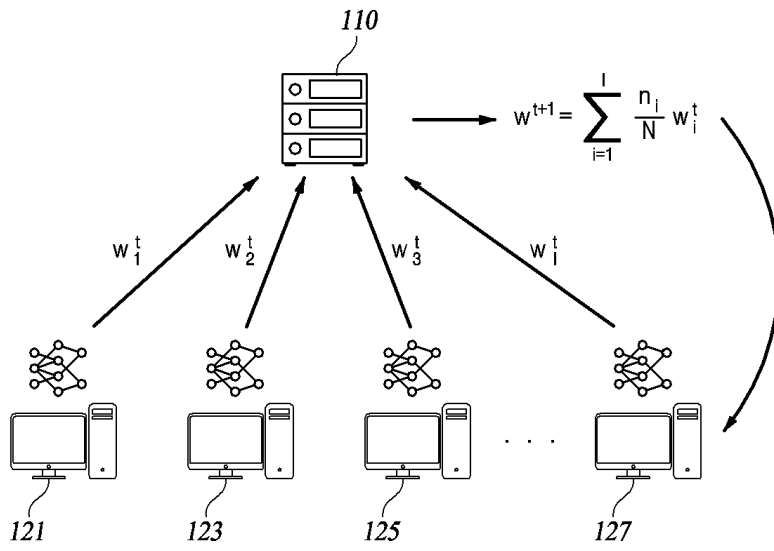
$\sum_{i=1}^I \frac{n_i}{N} h_{\theta}(g_{\theta}(w_i^t))$ 는 트레이닝 전역 모델, w^{t+1} 와 $\frac{n_i}{N} w_i^t$ 는 트레이닝 전역 모델의 레이블, w_i^t 는 제i 트레이닝 지역 모델, $h_{\theta}(g_{\theta}(w_i^t))$ 는 제i 디코딩 결과, $\frac{n_i}{N}$ 는 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치, N은 트레이닝 지역 모델들에 훈련된 데이터 샘플들의 총 개수, n_i 는 제i 트레이닝 지역 모델에 훈련된 데이터 샘플들의 개수를 나타낸다. 훈련 장치는 수학적 식 4에 따른 전역 모델 기반 손실함수가 0에 가까워지도록 인코더(310) 및 디코더(320)의 파라미터들을 업데이트할 수 있다.

- [0074] 이처럼, 오토인코더는 트레이닝 지역 모델들을 복원하는 것을 훈련하는 것이 아니라, 트레이닝 지역 모델들에 따른 전역 모델을 복원하는 데 이용되는 디코딩 결과들을 생성하도록 훈련된다. 디코딩 결과들에는 트레이닝 전역 모델의 레이블에 대한 컨텍스트 정보가 포함될 수 있다.
- [0075] 또한, 오토인코더는 다양한 데이터 분포를 갖는 트레이닝 지역 모델들을 모두 훈련하므로, 특정 데이터 분포에 과적합될 확률이 낮다.
- [0076] 이러한 훈련된 오토인코더를 적용하는 연합학습 시스템에서는, 전역 모델의 최적화 속도가 빨라지고, 전역 모델의 정확성 성능도 개선된다. 또한, 지역 모델의 압축에 따른 손실이 줄어들 수 있다.
- [0077] 도 4는 본 발명의 일 실시예에 따른 연합학습의 순서도이다.
- [0078] 도 4를 참조하면, 연합학습 시스템의 서버는 클라이언트들로부터 복수의 압축된 지역 모델들을 수신한다(S410).
- [0079] 클라이언트들은 사용자들의 훈련 데이터를 수집하고, 훈련 데이터를 이용하여 지역 모델들을 훈련하며, 인코더를 이용하여 지역 모델들을 압축한다. 이를 위해, 클라이언트들은 동일한 인코더들을 포함하며, 각 인코더는 각 클라이언트의 지역 모델을 압축한다.
- [0080] 서버는 디코더를 이용하여 압축된 지역 모델들로부터 디코딩 결과들을 획득한다(S420).
- [0081] 서버는 압축된 지역 모델들을 디코더에 입력하고, 디코더가 압축된 지역 모델들을 디코딩한 결과들을 획득한다.
- [0082] 여기서, 인코더들과 디코더가 지역 모델 기반 손실함수에 따라 훈련된 경우, 디코딩 결과는 압축된 지역 모델의 복원 결과이다. 반면, 인코더와 디코더가 전역 모델 기반 손실함수에 따라 훈련된 경우, 디코더의 디코딩 결과는 인코더에 입력되는 지역 모델과 다르다.
- [0083] 서버는 디코딩 결과들을 기반으로 전역 모델을 생성한다(S430).
- [0084] 서버는 디코딩 결과들의 평균 또는 가중평균을 전역 모델로 생성할 수 있다. 이때, 서버는 각 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 디코딩 결과를 가중하고, 가중된 디코딩 결과들을 평균하여 전역 모델을 생성할 수 있다.
- [0085] 서버는 전역 모델을 클라이언트들에게 전송한다(S440).
- [0086] 각 클라이언트는 전역 모델을 지역 모델로 설정하고, 훈련 데이터를 이용하여 지역 모델을 업데이트한다.
- [0087] 본 발명의 일 실시예에 따르면, 연합학습 시스템에서 클라이언트들이 동일한 인코더의 파라미터들을 공유하므로, 클라이언트의 수가 증가하더라도 연합학습이 수행될 수 있다. 또한, 클라이언트들의 인코더들이 서로 다른 데이터 분포에 과적합되어, 연합학습의 성능이 저하되는 것을 막을 수 있다.
- [0088] 도 5는 본 발명의 일 실시예에 따른 오토인코더의 훈련 방법의 순서도이다.
- [0089] 도 5를 참조하면, 훈련 장치는 서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 인코더를 이용하여 압축한다(S510).
- [0090] 훈련 장치는 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득한다(S520).
- [0091] 훈련 장치는 디코딩 결과들을 기반으로 트레이닝 전역 모델을 생성한다(S530).
- [0092] 구체적으로, 훈련 장치는 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 디코딩 결과를 가중하고, 가중된 트레이닝 디코딩 결과들을 평균하여 트레이닝 전역 모델을 생성한다.
- [0093] 훈련 장치는 트레이닝 지역 모델들을 기반으로 트레이닝 전역 모델의 레이블을 생성한다(S540).

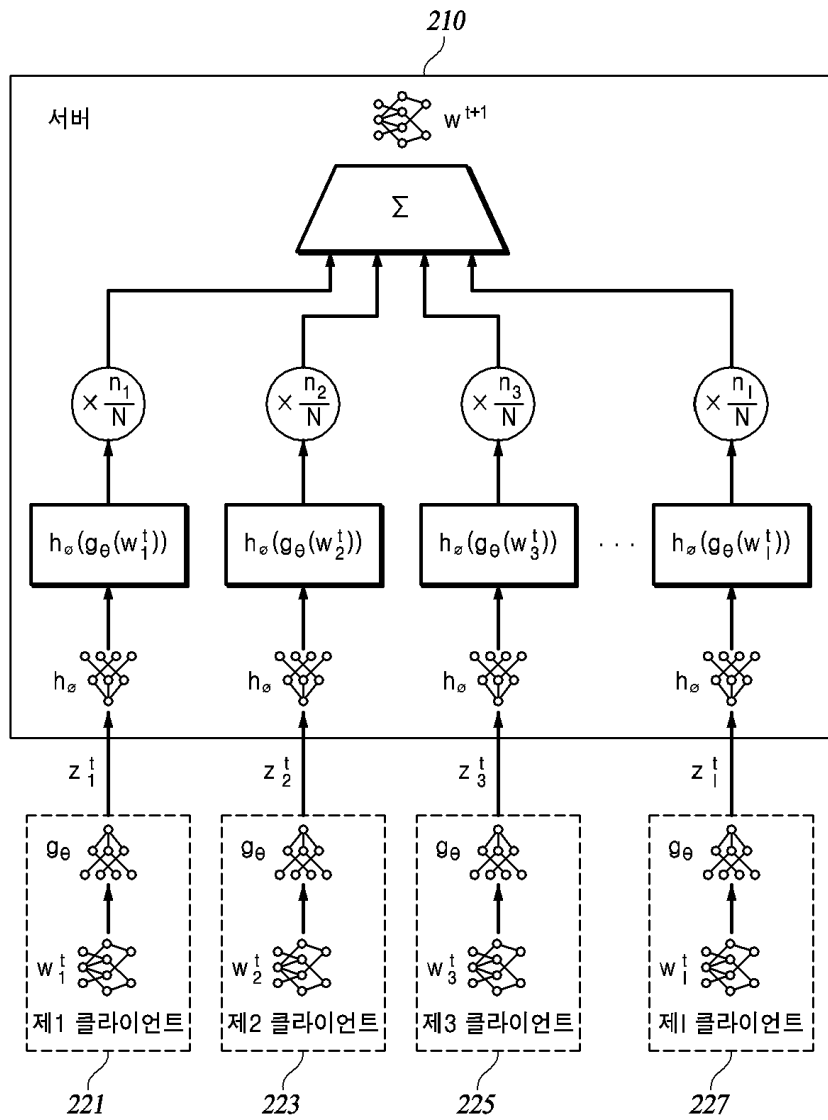
- [0094] 구체적으로, 훈련 장치는 각 트레이닝 지역 모델에 훈련된 데이터 샘플들의 수를 나타내는 가중치로 각 트레이닝 지역 모델을 가중하고, 가중된 트레이닝 지역 모델들을 평균하여 트레이닝 전역 모델의 레이블을 생성한다.
- [0095] 훈련 장치는 트레이닝 전역 모델 및 트레이닝 전역 모델의 레이블을 기반으로 인코더 및 디코더를 업데이트한다(S550).
- [0096] 특히, 훈련 장치는 트레이닝 전역 모델 및 트레이닝 전역 모델의 레이블 간 차이가 감소하도록 인코더 및 디코더를 업데이트한다.
- [0097] 전술한 과정들을 통해, 오토인코더는 단순히 지역 모델을 압축 및 복원하는 것이 아니라, 전역 모델을 위한 가중평균을 훈련함으로써, 연합학습 시스템의 훈련 속도를 높이고 통신 자원 소모량을 줄일 수 있다.
- [0098] 한편, 다른 실시예에 따르면, 훈련 장치는 오토인코더를 지역 모델 기반 손실함수를 이용하여 훈련할 수 있다. 구체적으로, 훈련 장치는 서로 다른 데이터 분포를 훈련한 복수의 트레이닝 지역 모델들을 인코더를 이용하여 압축하고, 디코더를 이용하여 압축된 트레이닝 지역 모델들로부터 트레이닝 디코딩 결과들을 획득한다. 이후, 훈련 장치는 트레이닝 지역 모델들과 트레이닝 디코딩 결과들 간 차이값들의 평균을 기반으로 인코더 및 디코더를 업데이트함으로써, 오토인코더를 훈련할 수 있다.
- [0099] 본 명세서에 설명되는 시스템들 및 기법들의 다양한 구현예들은, 디지털 전자 회로, 집적회로, FPGA(field programmable gate array), ASIC(application specific integrated circuit), 컴퓨터 하드웨어, 펌웨어, 소프트웨어, 및/또는 이들의 조합으로 실현될 수 있다. 이러한 다양한 구현예들은 프로그래밍가능 시스템 상에서 실행 가능한 하나 이상의 컴퓨터 프로그램들로 구현되는 것을 포함할 수 있다. 프로그래밍가능 시스템은, 저장 시스템, 적어도 하나의 입력 디바이스, 그리고 적어도 하나의 출력 디바이스로부터 데이터 및 명령들을 수신하고 이들에게 데이터 및 명령들을 전송하도록 결합되는 적어도 하나의 프로그래밍가능 프로세서(이것은 특수 목적 프로세서일 수 있거나 혹은 범용 프로세서일 수 있음)를 포함한다. 컴퓨터 프로그램들(이것은 또한 프로그램들, 소프트웨어, 소프트웨어 애플리케이션들 혹은 코드로서 알려져 있음)은 프로그래밍가능 프로세서에 대한 명령어들을 포함하며 "컴퓨터가 읽을 수 있는 기록매체"에 저장된다.
- [0100] 컴퓨터가 읽을 수 있는 기록매체는, 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 이러한 컴퓨터가 읽을 수 있는 기록매체는 ROM, CD-ROM, 자기 테이프, 플로피디스크, 메모리 카드, 하드 디스크, 광자기 디스크, 스토리지 디바이스 등의 비휘발성(non-volatile) 또는 비일시적인(non-transitory) 매체일 수 있으며, 또한 데이터 전송 매체(data transmission medium)와 같은 일시적인(transitory) 매체를 더 포함할 수도 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어, 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다.
- [0101] 본 명세서의 흐름도/타이밍도에서는 각 과정들을 순차적으로 실행하는 것으로 기재하고 있으나, 이는 본 개시의 일 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것이다. 다시 말해, 본 개시의 일 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 개시의 일 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 흐름도/타이밍도에 기재된 순서를 변경하여 실행하거나 각 과정들 중 하나 이상의 과정을 병렬적으로 실행하는 것으로 다양하게 수정 및 변형하여 적용 가능할 것이므로, 흐름도/타이밍도는 시계열적인 순서로 한정되는 것은 아니다.
- [0102] 이상의 설명은 본 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 실시예들은 본 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

도면

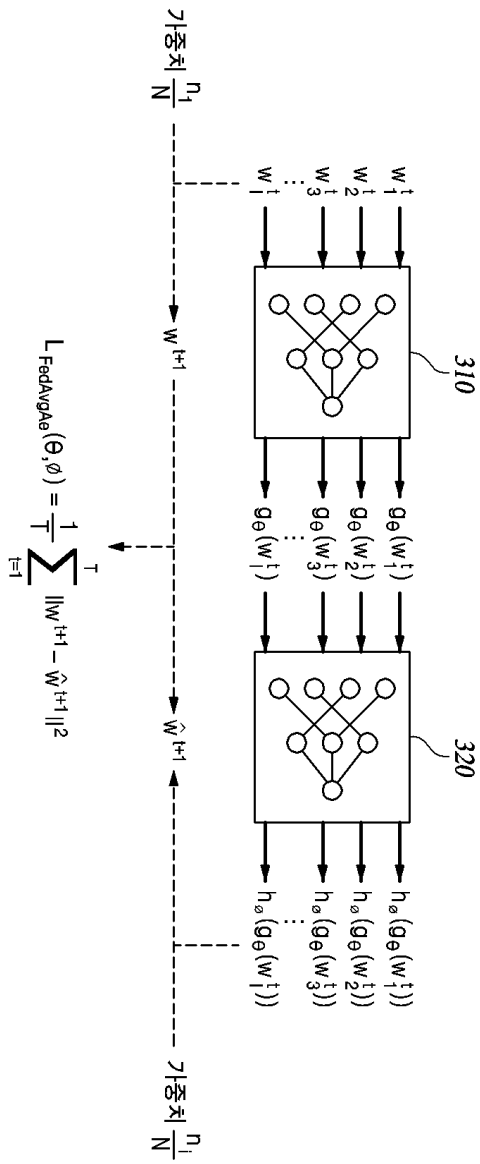
도면1



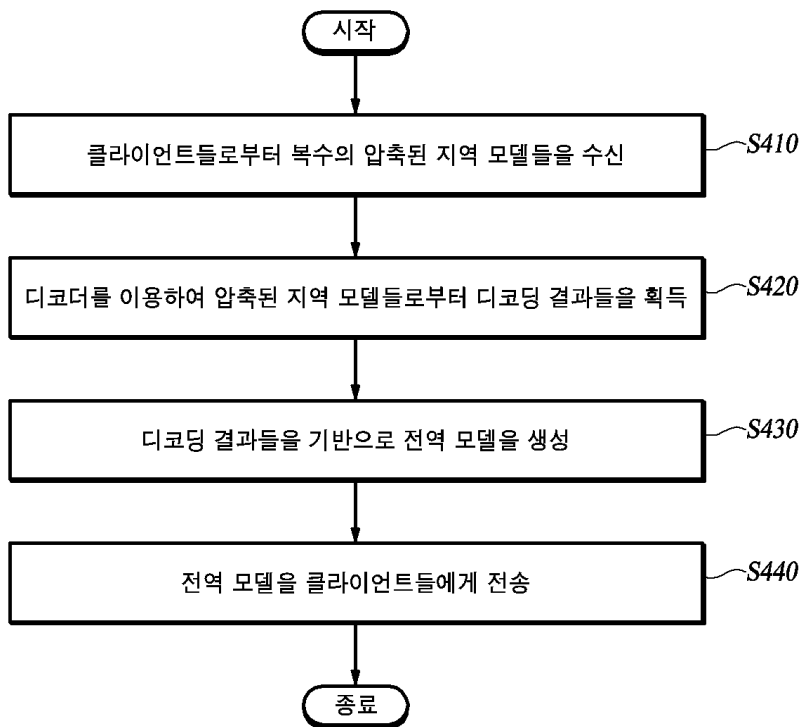
도면2



도면3



도면4



도면5

