



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0094089
(43) 공개일자 2024년06월25일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2023.01) G06N 3/04 (2023.01)
(52) CPC특허분류
G06N 3/084 (2023.01)
G06N 3/0454 (2013.01)
(21) 출원번호 10-2022-0159591
(22) 출원일자 2022년11월24일
심사청구일자 2022년11월24일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
김광인
울산광역시 울주군 언양읍 유니스트길 50
(74) 대리인
특허법인 무한

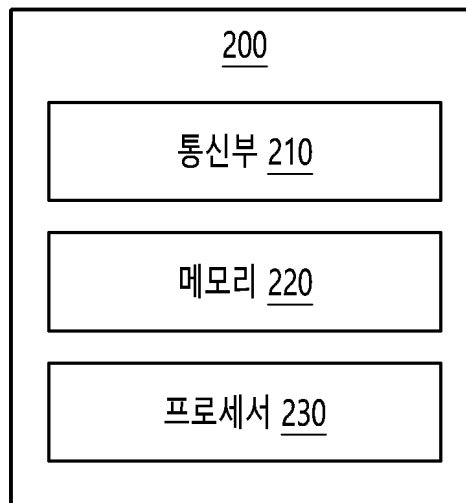
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법 및 장치

(57) 요약

일 실시예에 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법은 복수의 클라이언트(client)들에게 딥러닝 모델(deep learning model)을 위한 현재 반복(current iteration)의 학습 파라미터(parameter)를 전송하는 단계, 상기 복수의 클라이언트들로부터 상기 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트(local gradient) 값을 수신 하는 단계, 이전 반복(previous iteration)에서 그래디언트 정보 에 기초하여 현재 반복을 위해 산출된 현재 반복의 병합 가중치(combination weight) 에 기초하여 상기 복수의 현재 반복의 로컬 그래디언트 값들을 병합함으로써 현재 반복의 병합 그래디언트 값을 산출하는 단계 및 상기 현재 반복의 병합 그래디언트 값에 기초하여 다음 반복(next iteration)의 학습 파라미터를 업데이트하는 단계를 포함할 수 있다.

대표도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	1711152566
과제번호	2021-0-00537-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	차세대인공지능핵심원천기술개발
연구과제명	자기지도 학습에 의한 시각적 상식으로 영상에서 보이지 않는 부분을 복원하는 기술
기 여 율	1/1
과제수행기관명	한국전자기술연구원
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

프로세서에 의해 수행되는 분산 학습 환경에서 로컬 데이터의 오염에 강건한 딥러닝 학습 방법에 있어서,
 복수의 클라이언트(client)들에게 딥러닝 모델(deep learning model)을 위한 현재 반복(current iteration)의 학습 파라미터(parameter)를 전송하는 단계;
 상기 복수의 클라이언트들로부터 상기 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트(local gradient) 값들을 수신하는 단계;
 이전 반복(previous iteration)에서 그래디언트 정보에 기초하여 산출된 현재 반복의 병합 가중치(combination weight)에 기초하여 상기 현재 반복의 로컬 그래디언트 값들을 병합함으로써 현재 반복의 병합 그래디언트 값을 산출하는 단계; 및
 상기 현재 반복의 병합 그래디언트 값에 기초하여 다음 반복(next iteration)의 학습 파라미터를 업데이트하는 단계
 를 포함하는 딥러닝 학습 방법.

청구항 2

제1항에 있어서,
 이전 반복의 병합 가중치, 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값을 포함하는 그래디언트 정보에 기초하여, 각 로컬 그래디언트에 대해 다음 반복의 병합 가중치를 결정하는 단계
 를 포함하는 딥러닝 학습 방법.

청구항 3

제2항에 있어서,
 상기 병합 가중치를 결정하는 단계는,
 상기 다음 반복의 병합 가중치 결정 시 상기 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값의 차에 대한 음의 지수함수를 이용하는 것을 특징으로 하는 딥러닝 학습 방법.

청구항 4

제2항에 있어서,
 상기 현재 반복의 병합 가중치는,
 이전 반복의 병합 그래디언트 값 및 이전 반복의 로컬 그래디언트 값의 차의 제곱에 반비례하는,
 딥러닝 학습 방법.

청구항 5

제1항에 있어서,

상기 각 로컬 그래디언트 값에 적용되는 상기 병합 가중치의 총 합은 1이 되는 것을 특징으로 하는 딥러닝 학습 방법.

청구항 6

제1항에 있어서,

딥러닝 모델 학습 전, 상기 로컬 그래디언트 값을 산출하기 위하여 상기 복수의 클라이언트에게 학습 데이터를 로컬 데이터로 분산하여 전송하는 단계

를 포함하는 딥러닝 학습 방법.

청구항 7

제1항에 있어서,

상기 학습 파라미터를 업데이트하는 단계는,

상기 현재 반복의 로컬 그래디언트들을 병합함으로써 획득된 상기 현재 반복의 병합 그래디언트 값에 기초하여, 현재 반복의 학습 파라미터로부터 다음 반복의 학습 파라미터를 결정하는 단계

를 포함하는 딥러닝 학습 방법.

청구항 8

제1항 내지 제7항 중 어느 한 항의 방법을 수행하기 위한 명령어를 포함하는 하나 이상의 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 기록 매체.

청구항 9

로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 장치에 있어서,

통신부;

딥러닝 모델을 학습시키는 인스트럭션들(instruction)이 저장된 메모리; 및

상기 메모리에 저장된 인스트럭션들을 실행하는 프로세서

를 포함하고,

상기 프로세서는,

상기 통신부를 통해 복수의 클라이언트들에게 딥러닝 모델을 위한 현재 반복의 학습 파라미터를 전송하고 상기 복수의 클라이언트들로부터 상기 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트 값들을 수신하고, 이전 반복에서 그래디언트 정보에 기초하여 산출된 현재 반복의 병합 가중치에 기초하여 상기 현재 반복의 로컬 그래디언트 값들을 병합함으로써 현재 반복의 병합 그래디언트 값을 산출하며, 상기 현재 반복의 병합 그래디언트 값에 기초하여 다음 반복의 학습 파라미터를 업데이트하는 딥러닝 학습 장치.

청구항 10

제9항에 있어서,

상기 프로세서는

이전 반복의 병합 가중치, 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값을 포함하는 그

래디언트 정보에 기초하여, 각 로컬 그래디언트에 대해 다음 반복의 병합 가중치를 결정하는 딥러닝 학습 장치.

청구항 11

제10항에 있어서,

상기 프로세서는

상기 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값의 차에 대한 음의 지수함수를 이용하여 상기 다음 반복의 병합 가중치를 결정하는 딥러닝 학습 장치.

청구항 12

제10항에 있어서,

상기 프로세서는

상기 이전 반복의 병합 그래디언트 값 및 이전 반복의 로컬 그래디언트 값의 차의 제곱에 대하여 반비례하는 상기 현재 반복의 병합 가중치를 산출하는 딥러닝 학습 장치.

청구항 13

제9항에 있어서,

상기 각 로컬 그래디언트 값에 적용되는 상기 병합 가중치의 총 합은 1이 되는 것을 특징으로 하는 딥러닝 학습 장치.

청구항 14

제9항에 있어서,

상기 프로세서는

딥러닝 모델 학습 전, 상기 로컬 그래디언트 값을 산출하기 위하여 상기 통신부를 통해 상기 복수의 클라이언트에게 학습 데이터를 로컬 데이터로 분산하여 전송하는 딥러닝 학습 장치.

청구항 15

제9항에 있어서,

상기 프로세서는

상기 현재 반복의 로컬 그래디언트들을 병합함으로써 획득된 상기 현재 반복의 병합 그래디언트 값에 기초하여, 현재 반복의 학습 파라미터로부터 다음 반복의 학습 파라미터를 결정하는 딥러닝 학습 장치.

발명의 설명

기술 분야

로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법 및 장치가 제공된다.

[0001]

배경 기술

- [0003] 머신러닝(Machine Learning)은 인공지능(Artificial Intelligence, AI)의 하위 분야로 알고리즘을 통해 데이터를 분석하여 스스로 학습한 후, 이를 기반으로 어떠한 판단이나 예측을 하는 것을 의미할 수 있다. "학습(training)"은 데이터에 사람이 생각하는 정답(label)을 매겨서 계산하고, 사람의 생각과 차이가 나는 오류를 줄여가는 방법으로 반복하여 수정하는 과정을 의미할 수 있다.
- [0004] 딥 러닝(Deep Learning)은 머신러닝의 한 방법으로, 알고리즘과 규칙을 적용하여 학습이 수행되는 머신러닝과 달리 인공신경망(Artificial Neural Network, ANN)을 사용하여 학습할 수 있다.
- [0005] 위에서 설명한 배경기술은 발명자가 본원의 개시 내용을 도출하는 과정에서 보유하거나 습득한 것으로서, 반드시 본 출원 전에 일반 공중에 공개된 공지기술이라고 할 수는 없다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 대한민국 공개 특허공보 10- 2022-0111857(2022년 08월 10일 공고)에는 이미지 적대적 공격에 대비하기 위한 이미지 학습 장치 및 방법이 제시된다.
- (특허문헌 0002) 대한민국 공개 특허공보 10-2022-0030635(2022년 03월 11일 공고)에는 적대적 사례에 강인한 심층 신경망 모델을 위한 입력 장치 및 방법이 제시된다.

발명의 내용

해결하려는 과제

- [0008] 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법은 분산된 데이터에 적대적 공격이 발생하여 데이터가 오염되는 경우에도 오염된 데이터의 비중을 작게하여 병합함에 따라 딥러닝 모델의 성능을 유지할 수 있다.
- [0009] 다만, 기술적 과제는 상술한 기술적 과제들로 한정되는 것은 아니며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

- [0011] 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법은 복수의 클라이언트(client)들에게 딥러닝 모델(deep learning model)을 위한 현재 반복 (current iteration)의 학습 파라미터(parameter)를 전송하는 단계, 상기 복수의 클라이언트들로부터 상기 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트(local gradient) 값들을 수신 하는 단계, 이전 반복(previous iteration)에서 그래디언트 정보에 기초하여 산출된 현재 반복의 병합 가중치(combination weight)에 기초하여 상기 현재 반복의 로컬 그래디언트 값들을 병합함으로써 현재 반복의 병합 그래디언트 값을 산출하는 단계 및 상기 현재 반복의 병합 그래디언트 값에 기초하여 다음 반복(next iteration)의 학습 파라미터를 업데이트하는 단계를 포함할 수 있다.
- [0012] 이전 반복의 병합 가중치, 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값을 포함하는 그래디언트 정보에 기초하여, 각 로컬 그래디언트에 대해 다음 반복의 병합 가중치를 결정하는 단계를 포함할 수 있다.
- [0013] 상기 병합 가중치를 결정하는 단계는, 상기 다음 반복의 병합 가중치 결정 시 상기 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값의 차에 대한 음의 지수함수를 이용할 수 있다.
- [0014] 상기 현재 반복의 병합 가중치는, 이전 반복의 병합 그래디언트 값 및 이전 반복의 로컬 그래디언트 값의 차의 제곱에 반비례할 수 있다.

- [0015] 상기 각 로컬 그래디언트 값에 적용되는 상기 병합 가중치의 총 합은 1이 될 수 있다.
- [0016] 딥러닝 모델 학습 전, 상기 로컬 그래디언트 값을 산출하기 위하여 상기 복수의 클라이언트에게 학습 데이터를 로컬 데이터로 분산하여 전송하는 단계를 포함할 수 있다.
- [0017] 상기 학습 파라미터를 업데이트하는 단계는, 상기 현재 반복의 로컬 그래디언트들을 병합함으로써 획득된 상기 현재 반복의 병합 그래디언트 값에 기초하여, 현재 반복의 학습 파라미터로부터 다음 반복의 학습 파라미터를 결정하는 단계를 포함할 수 있다.
- [0018] 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 장치는 통신부, 딥러닝 모델을 학습시키는 인스트럭션들(instruction)이 저장된 메모리 및 상기 메모리에 저장된 인스트럭션들을 실행하는 프로세서를 포함하고, 상기 프로세서는, 상기 통신부를 통해 복수의 클라이언트들에게 딥러닝 모델을 위한 현재 반복의 학습 파라미터를 전송하고 상기 복수의 클라이언트들로부터 상기 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트 값들을 수신하고, 이전 반복에서 그래디언트 정보에 기초하여 산출된 현재 반복의 병합 가중치에 기초하여 상기 현재 반복의 로컬 그래디언트 값들을 병합함으로써 현재 반복의 병합 그래디언트 값을 산출하며, 상기 현재 반복의 병합 그래디언트 값에 기초하여 다음 반복의 학습 파라미터를 업데이트할 수 있다.
- [0019] 상기 프로세서는 이전 반복의 병합 가중치, 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값을 포함하는 그래디언트 정보에 기초하여, 각 로컬 그래디언트에 대해 다음 반복의 병합 가중치를 결정할 수 있다.
- [0020] 상기 프로세서는 상기 이전 반복의 로컬 그래디언트 값 및 이전 반복의 병합 그래디언트 값의 차에 대한 음의 지수함수를 이용하여 상기 다음 반복의 병합 가중치를 결정할 수 있다.
- [0021] 상기 프로세서는 상기 이전 반복의 병합 그래디언트 값 및 이전 반복의 로컬 그래디언트 값의 차의 제공에 대하여 반비례하는 상기 현재 반복의 병합 가중치를 산출할 수 있다.
- [0022] 상기 각 로컬 그래디언트 값에 적용되는 상기 병합 가중치의 총 합은 1이 될 수 있다.
- [0023] 상기 프로세서는 딥러닝 모델 학습 전, 상기 로컬 그래디언트 값을 산출하기 위하여 상기 통신부를 통해 상기 복수의 클라이언트에게 학습 데이터를 로컬 데이터로 분산하여 전송할 수 있다.
- [0024] 상기 프로세서는 상기 현재 반복의 로컬 그래디언트들을 병합함으로써 획득된 상기 현재 반복의 병합 그래디언트 값에 기초하여, 현재 반복의 학습 파라미터로부터 다음 반복의 학습 파라미터를 결정할 수 있다.
- [0025] 일 실시예에 따른 하나 이상의 컴퓨터 프로그램을 저장한 컴퓨터 판독 가능 기록 매체는, 방법을 수행하기 위한 명령어를 포함할 수 있다.

도면의 간단한 설명

- [0027] 도 1은 중앙 학습(centralized learning) 환경 및 분산 학습(distributed learning) 환경을 나타낸 예시를 도시한다.
- 도 2는 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 장치의 블록도를 도시한다.
- 도 3은 일 실시예에 따른 딥러닝 학습 장치가 적용된 분산 학습 시스템을 나타낸 예시를 도시한다.
- 도 4는 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법의 흐름도를 도시한다.
- 도 5는 일 실시예에 따른 반복시기에 따라 산출되는 값에 대한 예시를 도시한다.
- 도 6은 일 실시예에 따른 딥러닝 학습 방법과 다른 학습 방법의 성능 평가를 나타낼 수 있다.
- 도 7은 일 실시예에 따른 딥러닝 학습 방법을 이용한 예와 다른 학습 방법을 이용한 예를 나타낼 수 있다.
- 도 8은 도7의 딥러닝 학습 시 사용한 병합 가중치의 분포를 나타낼 수 있다.

발명을 실시하기 위한 구체적인 내용

- [0028] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0029] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0030] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0031] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0032] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0033] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.
- [0034] 도 1은 중앙 학습(centralized learning) 환경 및 분산 학습(distributed learning) 환경을 나타낸 예시를 도시한다.
- [0035] 중앙 학습(centralized learning) 환경(110)은 하나의 서버(111)에 모든 학습데이터가 저장되고 하나의 프로세서(112)에서 학습을 위한 프로세스를 수행할 수 있다. 프로세서(112)는 서버(111)에 있는 학습데이터를 이용하여 입력 도메인 X 를 출력 도메인 Y 로 매핑하는 함수 f 에 대하여 학습할 수 있다. 출력 도메인 Y 는 학습 문제의 특성에 따라 클래스 집합(예: 분류(classification) 문제) 또는 연속적인(continuous) 값(예: 회귀(regression) 문제)을 가질 수 있다.
- [0036] 학습데이터는 입력과 출력의 쌍으로 구성된 $Z = \{(x_1, y_1), \dots, (x_N, y_N)\} \subset X \times Y$ 가 제공되고, 이 경우 함수 f 는 하기의 수학적식을 이용하여 산출되는 에너지를 최소화하여 산출할 수 있다.

수학적식 1

$$\mathcal{E}(\mathbf{w}; Z) = \sum_{j=1}^N l(f(x_j), y_j)$$

- [0037]
- [0038] 전술한 수학적식 1에서 $\mathbf{w}$ 는 f 의 파라미터(parameter)를 나타내고, l 은 로스(loss) 함수를 나타낼 수 있다. 학습 문제 특성이 상술한 분류 문제인 경우, 로스는 하기의 수학적식2(예: 교차 엔트로피 오차(Cross Entropy Error, CEE))를 이용하여 산출될 수 있다.

수학식 2

$$l(a,b) = - \sum_{j=1}^c [a]_j \log ([b]_j)$$

[0039]

[0040]

전술한 수학식 2에서 $[a]_j$ 는 정답 레이블인 벡터 a 의 j 번째 요소를 나타낼 수 있고, 정답에 해당하는 인덱스의 원소만 1 값을 가지고 나머지는 0을 가진다. 또한, c 는 클래스(class)의 개수를 나타내고, $[b]_j$ 는 신경망 모델의 출력을 나타내며, $\log$ 는 밑이 e 인 자연로그를 나타낼 수 있다. 결과적으로 로스는 정답일때의 모델 값에 자연로그를 계산하여 산출될 수 있다.

[0041]

전술한 바와 같이 함수 f 는 수학식1의 에너지 $\mathcal{E}$ 를 최소화하여 얻을 수 있다. 이에 에너지 $\mathcal{E}$ 의 최소화는 경사하강(gradient descent) 기법을 통해 이루어질 수 있다. 경사하강(gradient descent) 기법은 에너지 $\mathcal{E}$ 가 최소가 되게 하는 파라미터를 찾는 방법으로 하기의 수학식3과 같이 반복적으로 $\mathbf{w}$ 를 업데이트할 수 있다.

수학식 3

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla_{\mathbf{w}} \mathcal{E}(\mathbf{w}; \mathbf{Z})$$

[0042]

[0043]

전술한 수학식 3에서 η 은 학습률(learning rate)을 나타내며 사용자가 직접 설정해야 하는 하이퍼 파라미터(Hyper parameter)일 수 있다. 또한, $\nabla_{\mathbf{w}} \mathcal{E}(\mathbf{w}; \mathbf{Z})$ 은 에너지 $\mathcal{E}$ 에 대한 파라미터 $\mathbf{w}$ 의 그래디언트(gradient) 벡터를 나타낼 수 있다.

[0044]

국방, 통신, 의료, IoT와 같은 산업계는 학습을 위한 데이터의 양이 대용량일 뿐만 아니라 정보 유출에 민감한 분야임에 따라 프라이버시(privacy)와 보안(security) 사항 또한 중요하다. 그러나 중앙 학습(centralized learning) 환경(110)은 단일 구성으로 처리 가능한 컴퓨팅 리소스가 제한될 수 있으며, 보안(security)과 프라이버시(privacy) 침해에 취약한 단점이 있다.

[0045]

이를 개선하기 위하여 분산 학습(distributed learning) 환경(120)은 데이터를 여러 클라이언트(client)에 분산시키고 학습 프로세스의 일부를 각 클라이언트 내에서 수행할 수 있다. 각 클라이언트 내에서 수행된 학습 프로세스 결과는 중앙 처리 장치(124)가 취합함에 따라 프로세스 처리에 대한 부하를 감소시킬 수 있으며, 중앙 처리 장치(124) 내 스토리지(storage) 또한 감소시킬 수 있다. 또한, 각 클라이언트에서 중앙 처리 장치(124)로 전송하는 데이터는 압축되어 전송됨에 따라 전송되는 데이터의 양이 작아 빠르게 송수신할 수 있다.

[0046]

분산 학습 환경(120)은 데이터 Z 가 K 개의 서로 다른 클라이언트 $\{\mathbf{C}_k\}_{k=1}^K$ 로 분산될 수 있다. 예시적으로, 제1 클라이언트(121)는 데이터 $\mathbf{z}_1$ 을 가질 수 있고, 제2 클라이언트(122)는 데이터 $\mathbf{z}_2$ 를 가질 수 있으며, 제3 클라이언트(123)는 데이터 $\mathbf{z}_3$ 을 가질 수 있다. 이에 분산 학습 환경(120)에서의 에너지 $\mathcal{E}$ 는 하기 수학식 4와 같이 표현될 수 있다. 또한, 파라미터를 업데이트하는 방법은 하기 수학식 5와 같이 표현될 수 있다.

수학식 4

$$\mathcal{E}(\mathbf{w}; \mathbf{Z}) = \sum_{k=1}^K \mathcal{E}(\mathbf{w}; \mathbf{Z}_k)$$

[0047]

수학식 5

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \sum_{k=1}^K \mathbf{g}_k(t)$$

[0048]

[0049]

전술한 수학식 5에서 $\mathbf{g}_k(t)$ 은 클라이언트 $\mathbf{C}_k$ 의 데이터 $\mathbf{Z}_k$ 에 대하여 계산된 로컬 그래디언트(local gradient) $\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}(t); \mathbf{Z}_k)$ 를 나타낼 수 있다.

[0050]

전술한 수학식 5를 통해 중앙 처리 장치(124)는 각 클라이언트에서 보낸 로컬 그래디언트를 병합하여 새로운 파라미터를 획득할 수 있다. 그러나 해킹 등의 적대적인 공격으로 인해 일부 클라이언트의 데이터는 오염될 수 있다. 이러한 오염된 데이터에 기초하여 산출된 로컬 그래디언트를 병합하여 학습하면 최종 신경망의 성능이 심각하게 하락할 수 있는 문제가 발생할 수 있다. 한편, 로컬화된 데이터의 세부 정보는 데이터 프라이버시 보호, 보안 강화 및 필요한 통신 비용을 감소시키는 사유 등으로 인해 일반적으로 클라이언트에 숨겨져 있을 수 있다. 이에 중앙 처리 장치(124)는 오염된 클라이언트의 ID를 파악할 수 없으며, 오염된 클라이언트의 개수 또한 미리 파악할 수 없다. 즉, 중앙 처리 장치(124)는 어떠한 클라이언트 데이터에서 오염이 발생하였는지와 얼마나 많은 클라이언트에서 오염이 발생하였는지를 확인하는 것이 어려울 수 있다. 전술한 데이터 오염 문제를 해결하기 위하여 사용자는 크게 학습데이터 자체를 정제하는 방식을 사용하거나, 오염된 데이터를 학습하여도 오분류를 하지 않는 강건한 모델이 되도록 반복적으로 재훈련하는 방식을 사용할 수 있다. 학습데이터 자체를 정제하는 방식은 오염 공격을 파악해야 하므로 데이터의 프라이버시 보호 및 보안 문제로 인해 적용이 어려울 수 있다. 이에 본 발명은 학습 시 오염된 데이터에 영향을 받지 않는 강건한 모델을 형성하는 방식을 응용하여 전술한 문제에 대한 해결 방안을 도출하였다.

[0051]

도 2는 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 장치의 블록도를 도시한다.

[0052]

일 실시예에 따른 장치(예: 딥러닝 학습 장치(200))는 통신부(210), 메모리(220) 및 프로세서(230)를 포함할 수 있다.

[0053]

통신부(210)는 복수의 클라이언트와 연결되어 데이터를 송수신할 수 있다. 복수의 클라이언트는 딥러닝 모델(deep learning model)을 학습할 수 있는 어떠한 기기도 가능할 수 있다. 복수의 클라이언트는, 예를 들면, 휴대폰, 스마트폰(smart phone), 노트북 컴퓨터(laptop computer), 디지털방송용 단말기, PDA(personal digital assistants), PMP(portable multimedia player), 네비게이션, 슬레이트 PC(slate PC), 하기 태블릿 PC(tablet PC), 울트라북(ultrabook), 웨어러블 디바이스(wearable device, 예를 들어, 위치형 단말기(smart watch), 글래스형 단말기(smart glass), HMD(head mounted display)) 등을 포함하는 정보통신기기와 멀티미디어기기 및 그에 대한 응용 기기일 수 있다.

[0054]

메모리(220)는 딥러닝 모델을 학습시키는 인스트럭션들(instruction) 및 모델 학습에 필요한 하이퍼 파라미터 값을 저장할 수 있다. 하이퍼 파라미터는 모델링할 때 사용자가 직접 세팅해주는 값으로 본 발명에서는 η, δ 및 σ^2 가 포함되나 이에 한정하지 않는다. 추가로, 메모리(220)는 모델 학습 전 복수의 클라이언트들에게 전송하는 로컬 데이터인 학습 데이터를 임시로 저장할 수 있다. 메모리(220)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 휘발성 저장장치일 수 있다. 예를 들면, 메모리(220)는 콤팩트 플래시(compact flash; CF) 카드, SD(secure digital) 카드, 메모리 스틱(memorystick), 솔리드 스테이트 드라이브(solid-state drive; SSD) 및 마이크로(micro) SD 카드 등과 같은 낸드 플래시 메모리(NAND flash memory), 하드 디스크 드라이브(hard disk drive; HDD) 등과 같은 마그네틱 컴퓨터 기억장치 및 CD-ROM, DVD-ROM 등과 같은 광학 디스크 드라이브(optical disc drive) 등을 포함할 수 있다.

[0055]

인스트럭션들은 프로세서(230)에서 수행하는 모든 동작들을 포함할 수 있다. 프로세서(230)는 메모리(220)에 저장된 인스트럭션들을 실행하여 딥러닝 모델을 학습할 수 있다. 프로세서(230)에 의해 수행되는 딥러닝 모델 학습은 도3을 들어 상세하게 설명한다.

[0056]

도 3은 일 실시예에 따른 딥러닝 학습 장치가 적용된 분산 학습 시스템을 나타낸 예시를 도시한다.

- [0057] (310)에서, 딥러닝 학습 장치(200)의 프로세서(230)는 통신부(210)를 통해 복수의 클라이언트들에게 딥러닝 모델을 위한 현재 반복(current iteration)의 학습 파라미터(311)(예: 신경망의 파라미터 벡터)를 전송할 수 있다. 반복은 최적의 학습 파라미터를 산출하기 위하여 반복되는 횟수를 의미할 수 있다.
- [0058] (320)에서, 딥러닝 학습 장치(200)의 프로세서(230)는 복수의 클라이언트들로부터 현재 반복의 학습 파라미터에 기초하여 연산된 현재 반복의 로컬 그래디언트 값들(321, 322, 323)을 수신할 수 있다. 만약 해킹 등과 같은 적대적 공격으로 인하여, 예를 들어, 로컬 데이터 z_3 가 오염되는 경우, 로컬 데이터 z_3 를 기초로 산출된 로컬 그래디언트 값(323) $\nabla_{\mathbf{w}} \mathcal{E}(\mathbf{w}; \mathbf{z}_3)$ 은 또한 잘못된 값일 수 있다. 이에 프로세서(230)는 현재 반복의 각 로컬 그래디언트 값들에 대응하여 현재 반복의 병합 가중치(combination weight)를 적용할 수 있다. 현재 반복의 병합 가중치는 이전 반복(previous iteration)에서 그래디언트 정보에 기초하여 산출된 값을 의미할 수 있다. 프로세서(230)는 현재 반복의 병합 가중치를 적용된 현재 반복의 로컬 그래디언트 값들을 병합하여 현재 반복의 병합 그래디언트 값을 산출할 수 있다. 이때 오염된 로컬 그래디언트 값 $\nabla_{\mathbf{w}} \mathcal{E}(\mathbf{w}; \mathbf{z}_3)$ 은 낮은 병합 가중치가 적용됨에 따라 병합 그래디언트 값 산출 시 기여도가 작을 수 있다. 이에 따라, 현재 반복의 병합 그래디언트 값에 기초하여 업데이트되는 다음 반복(next iteration)의 학습 파라미터는 오염된 데이터의 영향을 적게 받을 수 있다. 프로세서(230)를 통해 결정되는 각 값과 반복에 대한 관계는 도5에서 후술한다.
- [0059] 프로세서(230)는 딥러닝 모델 학습 전 통신부(210)를 통해 복수의 클라이언트들에게 메모리(220)에 저장된 학습 데이터를 로컬 데이터로 분산하여 전송할 수 있다. 각각의 클라이언트들은 수신된 각각의 로컬 데이터가 상호 중복되지 않을 수 있다. 이에 따라 각각의 클라이언트들은 동일한 학습 파라미터(311)를 수신하여도 각기 다른 로컬 그래디언트 값(321, 322, 323)을 산출할 수 있다. 한편, 적대적 공격에 의한 오염은 이렇게 분산된 로컬 데이터에서 발생할 수 있다.
- [0060] 이에 프로세서(230)는 전술한 바와 같이 병합 가중치를 로컬 그래디언트 값에 적용할 수 있다. 프로세서(230)는 하기 수학적 식 6의 연산이 구현되어 병합 가중치를 결정할 수 있다.

수학적 식 6

[0061]
$$\alpha_k(t+1) = (1 - \delta)\alpha_k(t-1) + \delta \overline{\alpha}_k(t)$$

- [0062] 전술한 수학적 식 6에서 δ 는 0에서 1사이의 값으로 각 반복마다 α_k 가 얼마나 빨리 변할 수 있는지를 조정할 수 있다. 예를 들어, δ 가 0에 가까운 경우, $\alpha_k(t)$ 는 이전 반복의 $\alpha_k(t-1)$ 와 유사하게 되어 시간이 지나도 그 값이 거의 변하지 않는 특성을 보일 수 있다. 이와 반대로 δ 가 1에 가까운 경우, α_k 는 각 반복마다 매우 빨리 변하게 되는 특성을 보일 수 있다. 이에 본 발명에서는 실험을 통해 δ 의 값을 예시적으로 0.1로 정하였으나, 이로 한정하는 것은 아니다. 또한, 전술한 수학적 식 6에서 $\overline{\alpha}_k(t)$ 는 하기 수학적 식 7을 통해 산출할 수 있다.

수학적 식 7

[0063]
$$\overline{\alpha}_k(t) = \exp\left(-\frac{\|\mathbf{g}_k(t-1) - \mathbf{q}_k(t-1)\|^2}{\sigma^2}\right)$$

- [0064] 전술한 수학적 식 6에서 σ^2 는 병합 가중치 α_k 가 얼마나 균등한지를 결정할 수 있다. σ^2 는 0보다 큰 특징을 가질 수 있다. 예를 들어, σ^2 이 매우 작은 경우, 적은 수의 로컬 그래디언트($\mathbf{g}_k$)에 대해서만 해당 $\{\alpha_k\}$ 값이 0에 비해 유의미하게 크게 되어 이들만 $\mathbf{q}_k$ 를 구성하는 기여할 수 있다. 이와 반대로 σ^2 이 매우 큰 경우, α_k 값이 모두 $\frac{1}{K}$ 에 가깝게 돼 모든 그래디언트가 동일하게 $\mathbf{q}_k$ 를 구성하는 기여할 수 있다. 이에 본 발명에서는 실험을 통해 σ^2 의 값을 예시적으로 0.2로 정하였으나, 이로 한정하는 것은 아니다.

[0065] 또한, $\overline{\alpha_k}$ 은 병합 그래디언트인 $\mathbf{q}_k$ 와 로컬 그래디언트인 $\mathbf{g}_k$ 간의 거리(차)의 제곱에 대한 음의 지수함수로 반비례하는 것을 볼 수 있다. 이에 따라 $\mathbf{q}_k$ 와 $\mathbf{g}_k$ 간의 거리가 멀면 $\overline{\alpha_k}$ 값은 작아지고, $\mathbf{q}_k$ 와 $\mathbf{g}_k$ 간의 거리가 가까우면 $\overline{\alpha_k}$ 값은 커질 수 있다. 또한, $\mathbf{q}_k$ 와 $\mathbf{g}_k$ 간의 거리가 먼 것은 해당 로컬 그래디언트가 오염된 것을 의미함에 따라 병합 시 작은 가중치 값에 의해 자동으로 영향이 작아질 수 있다.

[0066] 프로세서(230)는 전술한 바와 같이 병합 가중치를 로컬 그래디언트 값에 적용하여 병합시킴으로써 병합 그래디언트 값을 획득할 수 있다. 프로세서(230)는 하기 수학적 식 8의 연산이 구현되어 병합 그래디언트 값을 결정할 수 있다.

수학적 식 8

[0067]
$$\mathbf{q}(t) = \sum_{k=1}^K \alpha_k(t) \mathbf{g}_k(t)$$

[0068] 전술한 수학적 식 8에서 병합 가중치 $\{\alpha_k(t)\}_{k=1}^K$ 들은 양의 값으로 그 합이 1이 될 수 있다. 그러나, 전술한 수학적 식 6을 통해 얻은 병합 가중치 값들은 각 반복에서 자동으로 그 합이 1이 되지 않을 수 있다. 이에 따라 최종 병합 가중치 값은 수학적 식 6에서 얻은 값들을 그들의 합으로 나눠서 얻을 수 있다.

[0069] 또한, $\mathbf{q}(t)$ 는 전술한 수학적 식 8을 통해 결정할 수 있으며, 하기 수학적 식 9에 의해서도 획득할 수 있다.

수학적 식 9

[0070]
$$\varepsilon^w(\mathbf{r}) = \frac{1}{2} \sum_{k=1}^S [\mathbf{c}(t)]_k \|\mathbf{h}_k(t-1) - \mathbf{r}\|^2$$

[0071] 전술한 수학적 식 9에서 k 는 1 이상 $S-1$ 이내일 수 있으며, $\mathbf{h}_k(t) = \mathbf{g}_k(t)$ 을 나타낼 수 있다. 또한, $\tilde{\mathbf{c}}(t)$ 는 $\frac{\mathbf{c}(t)}{\|\mathbf{c}(t)\|}$ 를 나타낼 수 있으며, $[\tilde{\mathbf{c}}(t)]_k$ 는 $\delta\alpha(t)_k$ 를 나타낼 수 있다. 만약 k 가 S 인 경우, $\mathbf{h}_k(t)$ 는 $\mathbf{q}(t)$ 를 나타내고, $[\tilde{\mathbf{c}}(t)]_k$ 는 $1 - \delta$ 을 나타내며, $[\mathbf{c}]_k$ 는 벡터 $\mathbf{c}$ 의 k 번째 요소일 수 있다. $\mathbf{r}$ 은 에너지 ε 를 최소화하는 최적화 변수로 산출되어야 하는 값일 수 있다. 이에 $\mathbf{h}_k(t-1) - \mathbf{r}$ 의 최소 값은 각각의 $\mathbf{h}_k(t-1)$ 값과 $\mathbf{r}$ 값이 상호 비슷한 값인 경우일 수 있다. 또한, $[\mathbf{c}(t)]_k$ 값이 0인 경우에는 $\mathbf{h}_k(t-1) - \mathbf{r}$ 이 중요하지 않고, $[\mathbf{c}(t)]_k$ 값이 1인 경우에는 $\mathbf{h}_k(t-1) - \mathbf{r}$ 이 중요할 수 있다. 이에 전술한 $\mathbf{q}(t)$ 의 산출 방식과 동일하게 병합 가중치는 이전 반복의 병합 그래디언트 값 및 이전 반복의 로컬 그래디언트 값의 차의 제곱에 반비례할 수 있다.

[0072] 프로세서(230)는 전술한 바와 같이 병합 그래디언트 값에 기초하여 학습 파라미터를 결정할 수 있다. 프로세서(230)는 하기 수학적 식 10의 연산이 구현되어 학습 파라미터를 결정할 수 있다.

수학적 식 10

[0073]
$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta K \mathbf{q}(t)$$

[0074] 만약 전술한 수학적 식 10에서 병합 가중치 $\{\alpha_k(t)\}_{k=1}^K$ 들의 합이 1이고, $\alpha_k(t)$ 또한 양의 값으로 모두 같은 값을 가지는 경우, $\alpha_k(t)$ 는 $\frac{1}{K}$ 이 될 수 있다. 이에 수학적 식 10은 전술한 수학적 식 5와 동일한 식일 수 있다.

[0075] 도 4는 일 실시예에 따른 로컬 데이터가 오염된 분산 학습 환경에서 오염에 강건한 딥러닝 학습 방법의 흐름도

를 도시한다.

- [0076] 단계(410)에서, 프로세서는 복수의 클라이언트들에게 딥러닝 모델을 위한 현재 반복(t)의 학습 파라미터를 전송할 수 있다. 현재 반복의 학습 파라미터 $w(t)$ 는 이전 반복(t-1)에서 산출한 학습 파라미터일 수 있다.
- [0077] 단계(420)에서, 프로세서는 복수의 클라이언트들로부터 현재 반복의 학습 파라미터 $w(t)$ 에 기초하여 연산된 현재 반복의 로컬 그래디언트 값 $g_k(t)$ 들을 수신할 수 있다. 프로세서는 딥러닝 모델 학습 전 클라이언트가 로컬 그래디언트 값을 산출할 수 있도록 학습 데이터를 로컬 데이터로 분산하여 전송할 수 있다.
- [0078] 단계(430)에서, 프로세서는 이전 반복(t-1)에서 그래디언트 정보에 기초하여 산출된 현재 반복(t)의 병합 가중치에 기초하여 복수의 현재 반복(t)의 로컬 그래디언트 값들을 병합함으로써 현재 반복(t)의 병합 그래디언트 값을 산출할 수 있다. 그래디언트 정보는 현재 반복(t)의 병합 가중치를 산출하기 위하여 이전 반복(t-1)에서 사용한 값들일 수 있다. 이에 현재(t) 반복의 병합 가중치를 산출하기 위하여 사용한 그래디언트 정보는 (t-3) 반복에서 산출한 (t-2) 반복의 병합 가중치, (t-2) 반복에서 산출한 (t-2) 반복의 로컬 그래디언트 값 및 (t-2) 반복에서 산출한 (t-2) 반복의 병합 그래디언트 값을 포함할 수 있다. 또한, 프로세서는 전송한 그래디언트 정보와 상이한 그래디언트 정보를 이용하여 현재(t) 반복에서 다음(t+1) 반복을 위한 다음(t+1) 반복의 병합 가중치를 산출할 수 있다. 이에 다음(t+1) 반복의 병합 가중치 산출을 위한 그래디언트 정보는 (t-2) 반복에서 산출한 이전 반복(t-1)의 병합 가중치, 이전(t-1) 반복에서 산출한 이전 반복(t-1)의 로컬 그래디언트 값 및 이전(t-1) 반복에서 산출한 이전 반복(t-1)의 병합 그래디언트 값을 포함할 수 있다. 한편, 프로세서는 최초 반복(t=1) 시 병합 가중치 $\alpha_k(1)$ 를 $\frac{1}{k}$ 로 초기화하여 수행할 수 있다. 이는 병합 가중치가 모두 균일한 것을 의미할 수 있다.
- [0079] 단계(440)에서, 프로세서는 현재 반복(t)의 병합 그래디언트 값에 기초하여 다음 반복(t+1)의 학습 파라미터를 업데이트할 수 있다.
- [0080] 본 발명은 수학식 3에서 전송한 바와 같이 반복적으로 학습 파라미터 w 를 업데이트하여 에너지 $\mathcal{E}$ 가 최소가 되어 획득할 수 있는 함수 f 의 학습을 수행할 수 있다. 이에 프로세서는 각 반복시기에 병합 그래디언트 값 및 병합 가중치를 산출하여 최종적으로 학습 파라미터를 산출할 수 있다.
- [0081] 도 5는 일 실시예에 따른 반복시기에 따라 산출되는 값에 대한 예시를 도시한다.
- [0082] 현재 반복(530)을 기준으로 프로세서는 현재 반복(530)의 로컬 그래디언트 값(531)을 산출하기 위하여 복수의 클라이언트들에게 이전 반복(520)의 학습 파라미터(524)를 전송할 수 있다. 복수의 클라이언트들은 수신된 학습 파라미터(524)를 기초로 현재 반복(530)의 로컬 그래디언트 값(531)을 산출할 수 있다.
- [0083] 또한 프로세서는 병합 그래디언트 값(532)을 산출하기 위하여 복수의 클라이언트들로부터 수신받은 각각의 로컬 그래디언트 값(531)에 이전 반복(520)에서 산출한 각각의 병합 가중치(523)를 적용하여 병합할 수 있다.
- [0084] 그 다음, 프로세서는 다음 반복(540)을 위한 병합 가중치(533)를 산출할 수 있으며, 그 순서는 다음 반복(540)의 학습 파라미터(534)를 산출하는 단계 전/후 다 가능할 수 있다. 프로세서는 t-2 반복 시기(510)에 산출한 이전 반복(520)의 병합 가중치(511), 이전 반복(520)의 로컬 그래디언트 값(521) 및 이전 반복(520)의 병합 그래디언트 값(522)에 기초하여 다음 반복(540)을 위한 병합 가중치(533)를 산출할 수 있다. 이전 반복의 값을 이용하여 산출된 병합 가중치는 반복이 진행됨에 따라 부드럽게(smooth) 변할 수 있다. 이에 부드러운 병합 가중치 변화는 반복에서 있을 오류의 영향을 최소화하거나 무시할 수 있도록 한다.
- [0085] 프로세서는 다음 반복(540)의 학습 파라미터(534)를 산출하기 위하여 이전 반복(520)의 학습 파라미터(524)와 현재 반복(530)의 병합 그래디언트 값(532)을 이용할 수 있다.
- [0086] 도 6은 일 실시예에 따른 딥러닝 학습 방법과 다른 학습 방법의 성능 평가를 나타낼 수 있다.
- [0087] 본 발명의 일 실시예에 따른 딥러닝 학습 방법의 평가는 Caltech256, CIFAR10과 CIFAR100, FashionMNIST, 및 CINIC10 데이터 베이스를 사용하여 평가하였다. 또한, 성능 비교를 위해 Federated averaging 방법(FedAvg), Pillutla 등의 그래디언트 병합 방법(GeoMed), Turan 등의 그래디언트 병합 방법(MedTh)의 성능 또한 평가하였다. 각 데이터는 100개의 클라이언트로 분산되었고, 각 클라이언트의 데이터는 10개 단위로 10개에서부터 60개 까지 오염시켰다. 본 발명의 딥러닝 학습 방법은 전체의 데이터 베이스에서 뛰어난 성능을 보여주었으며, 오염도가 강할수록 다른 방법에 비해 더욱 높은 성능을 보여주었다.

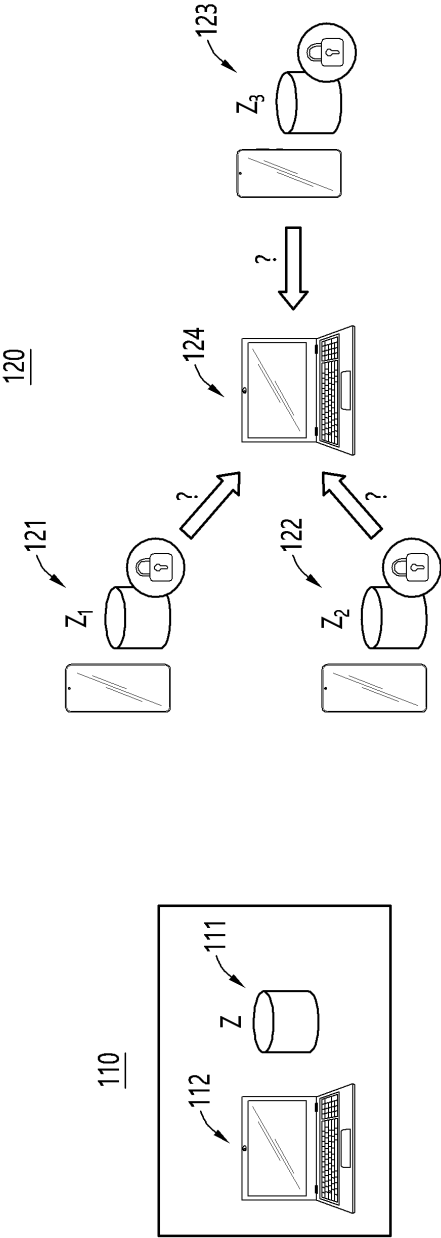
- [0088] 도 7은 일 실시예에 따른 딥러닝 학습 방법을 이용한 예와 다른 학습 방법을 이용한 예를 나타낼 수 있다.
- [0089] 도 8은 도7의 딥러닝 학습 시 사용한 병합 가중치의 분포를 나타낼 수 있다.
- [0090] 도 7(a)는 ground-truth로 학습하고자 하는 데이터의 실제 값(정답)이고, 도 7(b)는 ground-truth로부터 노이즈 샘플로 생성한 학습 세트(training set)일 수 있다. 이러한 학습 세트는, 예를 들어, 125개의 클라이언트들에게 배포될 수 있다. 각각의 학습 반복에서, 예를 들어, 40개의 클라이언트들의 그래디언트는 적대적 공격을 시뮬레이션하는 가우스 랜덤 벡터로 대체되었다. 이 경우, FedAvg 방법은 도 7(c)에서 묘사된 것과 같이 도 7(a)의 ground-truth에 비하여 상이한 형태를 묘사한다. 반면에 본 발명의 일 실시예에 따른 딥러닝 학습 방법은 도 7(d)에서 묘사된 것과 같이 도 7(a)의 ground-truth와 거의 유사한 형태를 묘사한다. 또한, 도 7(c)의 FedAvg 방법의 형태보다 정확한 형태를 이루고 있다. 정확도에 있어서도 FedAvg 방법은 81.38%의 수행 능력을 보이나, 본 발명의 딥러닝 학습 방법은 90.81%의 수행 능력을 보여주었다. 도8에서 나타낸 바와 같이 본 발명의 일 실시예에 따른 딥러닝 학습 방법은 오염된 클라이언트에 낮은 병합 가중치(810)를 부여하고 오염되지 않은 클라이언트에게는 높은 병합 가중치(820)를 부여한다. 이에 낮은 병합 가중치를 가진 오염된 그래디언트 값의 영향은 억제하여 학습에 영향을 미치지 않게 할 수 있다.
- [0091] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0092] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0093] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0094] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.
- [0095] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수

있다.

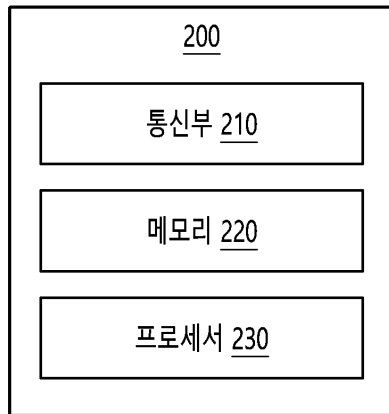
[0096] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

도면1

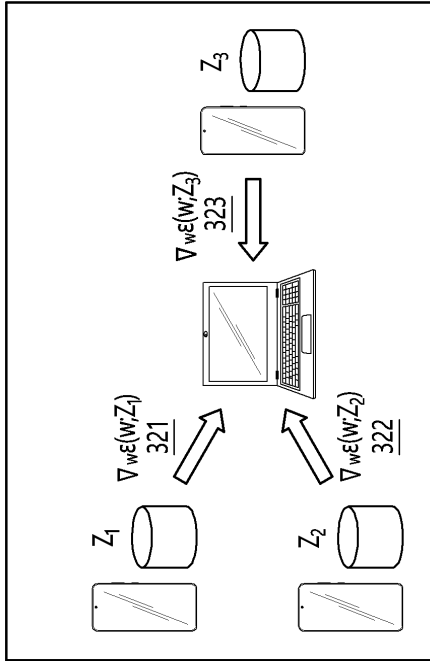


도면2

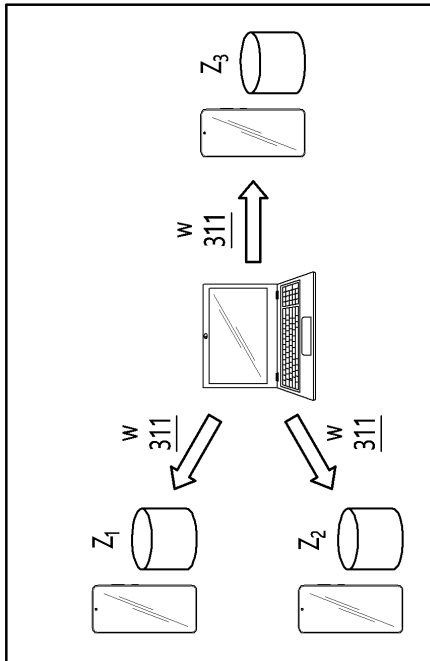


도면3

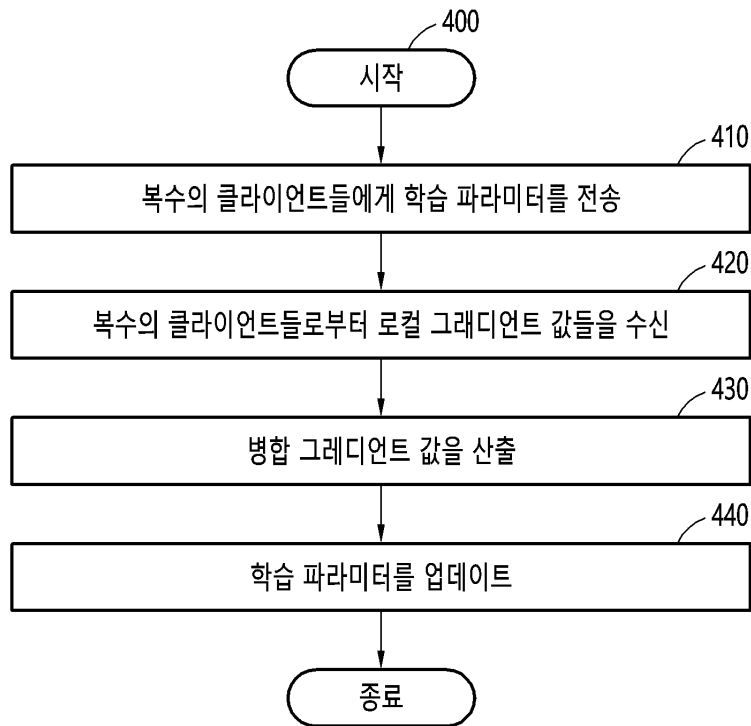
320



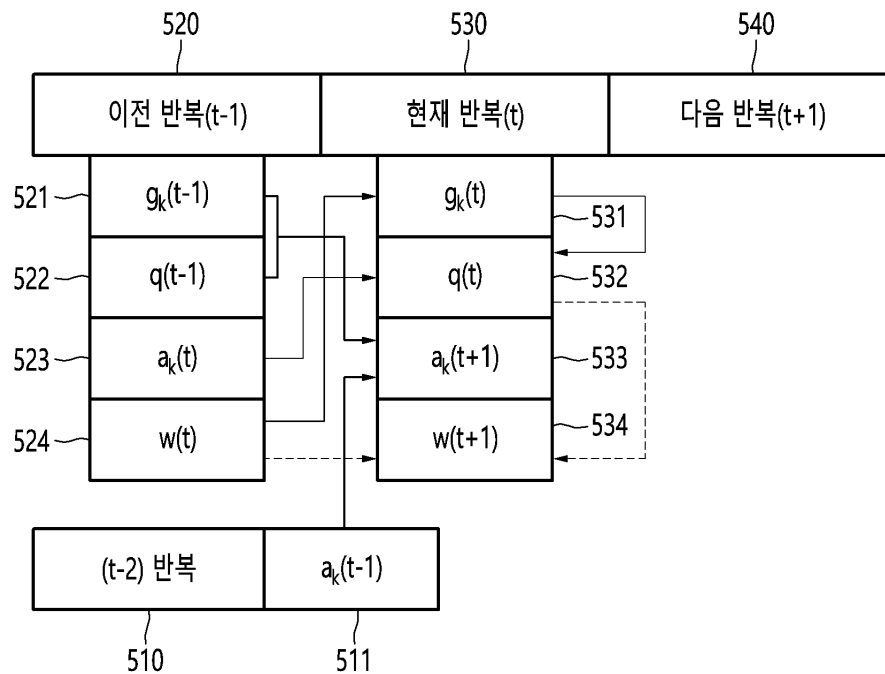
310



도면4



도면5



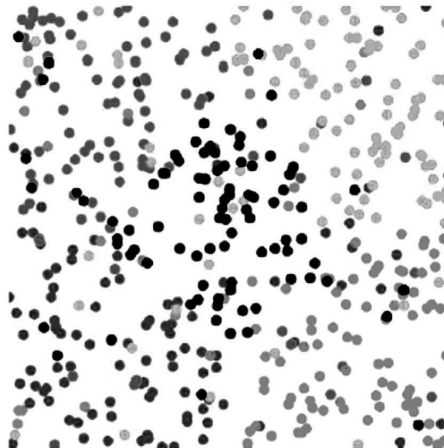
도면6

Dataset	Method	Accuracy (%)					
오염된 클라이언트 수		10	20	30	40	50	60
Caltech256	FedAvg	76.23 (0.51)	75.28 (0.13)	74.65 (0.38)	72.91 (0.44)	71.37 (0.51)	67.12 (0.54)
	GeoMed	76.57 (0.28)	75.59 (0.37)	74.72 (0.40)	72.59 (0.37)	67.49 (0.47)	59.27 (0.66)
	MedTh	76.06 (0.30)	75.69 (0.29)	74.81 (0.42)	73.04 (0.80)	68.86 (0.82)	44.66 (4.49)
	Ours	76.10 (0.14)	75.54 (0.31)	75.01 (0.27)	74.15 (0.29)	72.92 (0.40)	70.71 (0.53)
CIFAR10	FedAvg	88.06 (0.42)	87.35 (0.25)	86.56 (0.15)	85.99 (0.33)	85.33 (0.30)	84.99 (0.38)
	GeoMed	88.21 (0.20)	88.15 (0.70)	87.87 (0.79)	87.90 (1.34)	85.88 (0.44)	82.51 (0.41)
	MedTh	90.09 (0.09)	89.97 (0.19)	89.73 (0.20)	88.65 (2.70)	79.60 (17.93)	31.72 (21.78)
	Ours	90.10 (0.20)	90.03 (0.19)	89.83 (0.17)	89.85 (0.17)	89.56 (0.20)	89.33 (0.18)
CIFAR100	FedAvg	67.26 (0.32)	66.19 (0.23)	64.87 (0.35)	63.41 (0.38)	61.69 (0.38)	60.38 (0.25)
	GeoMed	67.84 (0.43)	67.61 (0.38)	66.60 (0.55)	65.56 (0.54)	57.89 (0.72)	48.32 (0.86)
	MedTh	67.63 (0.23)	67.13 (0.42)	66.55 (0.25)	64.81 (0.52)	39.50 (2.07)	36.82 (1.69)
	Ours	67.57 (0.25)	67.15 (0.29)	66.63 (0.14)	65.81 (0.22)	65.23 (0.25)	63.66 (0.64)
FashionMNIST	FedAvg	86.48 (0.19)	83.12 (0.24)	81.93 (0.35)	81.30 (0.31)	80.80 (0.50)	80.47 (0.26)
	GeoMed	83.29 (1.65)	82.37 (0.40)	82.42 (1.72)	81.20 (0.36)	80.41 (0.45)	78.37 (0.30)
	MedTh	86.75 (0.11)	86.55 (0.22)	86.27 (0.22)	83.57 (4.56)	76.16 (14.07)	52.65 (23.84)
	Ours	87.01 (0.23)	86.73 (0.16)	86.63 (0.27)	86.40 (0.51)	86.24 (0.16)	86.41 (0.27)
CINIC10	FedAvg	78.20 (0.11)	76.88 (0.21)	76.04 (0.16)	75.41 (0.29)	74.81 (0.27)	74.38 (0.13)
	GeoMed	78.13 (1.95)	79.01 (1.86)	77.49 (2.61)	78.92 (1.20)	75.64 (1.28)	71.51 (0.75)
	MedTh	79.48 (0.11)	79.31 (0.15)	79.06 (0.10)	78.57 (0.15)	57.10 (23.87)	42.19 (20.96)
	Ours	80.03 (0.13)	79.95 (0.17)	79.85 (0.16)	79.64 (0.32)	79.19 (0.50)	78.69 (0.52)

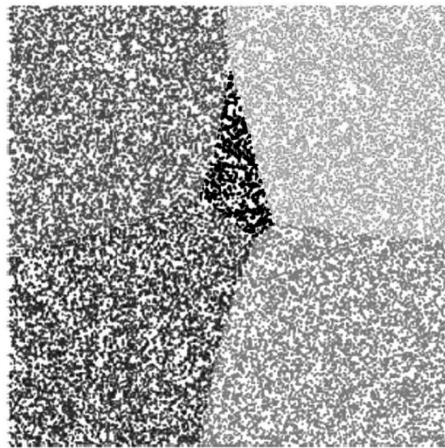
도면7



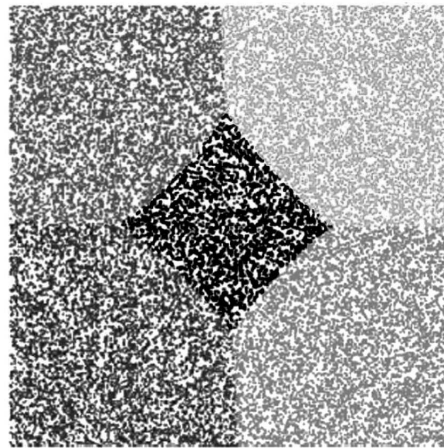
(a) Ground truth



(b) Training data



(c) FedAvg



(d) Ours

도면8

