(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0047229
(43) 공개일자 2023년04월07일

(51) 국제특허분류(Int. Cl.)
G16H 50/70 (2018.01) *G06N 3/04* (2023.01)
G06N 3/08 (2023.01) *G16H 50/20* (2018.01)
G16H 50/50 (2018.01) *G16H 70/00* (2018.01)
(52) CPC특허분류
G16H 50/70 (2018.01)
G06N 3/042 (2023.01)
(21) 출원번호 10-2021-0129372
(22) 출원일자 2021년09월30일
심사청구일자 2021년09월30일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
단국대학교 산학협력단
경기도 용인시 수지구 죽전로 152(죽전동, 단국대학교내)
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
(72) 발명자
김중헌
서울특별시 동작구 상도로 346-2, 212동201호(상도동, 상도엠코타운 애스톤파크)
윤형진
서울특별시 강남구 압구정로 151,123동 804호 (압구정동, 현대아파트)
(뒷면에 계속)
(74) 대리인
윤귀상

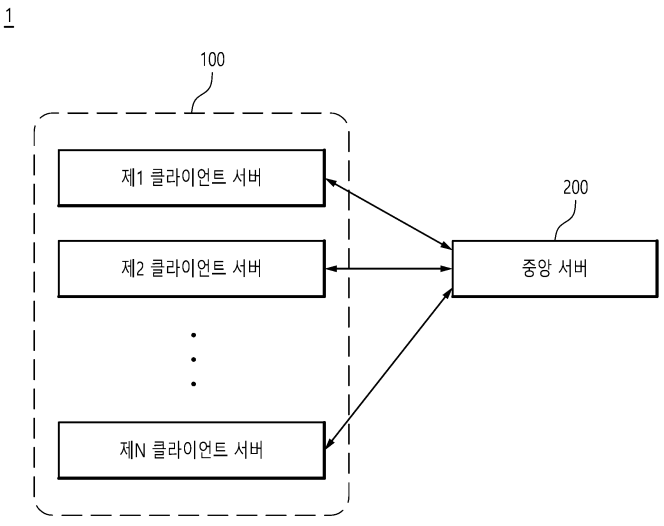
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **의료데이터 스플릿 학습 시스템과 이의 제어방법 및 그 방법을 수행하기 위한 기록매체**

(57) 요 약

병명을 가진 상기 의료데이터를 요청하는 중앙 서버; 및 중앙 서버로부터 수신한 의료데이터 요청신호에 따라 병명을 가진 원본 의료데이터를 선택하고, 선택된 원본 의료데이터의 특징맵을 추출하여 적어도 하나 이상의 훈련 의료데이터를 생성하는 의료기관 측에 마련된 적어도 하나 이상의 클라이언트 서버;를 포함하는 의료데이터 스플릿 학습 시스템을 제공한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2023.01)
G16H 50/20 (2018.01)
G16H 50/50 (2018.01)
G16H 70/00 (2021.08)

(72) 발명자
하유정
서울특별시 강남구 삼성로 417,103동 1601호 (대치동, 대치포스코더샵아파트)
유민재
경기도 성남시 분당구 탄천로 35,508-2203 (이매동, 아름마을풍림아파트)
이구상
서울특별시 성북구 개운사길 21-3,502호(안암동5가)
정소이
경기도 수원시 팔달구 중부대로223번길 54-13,202동 501호(우만동, 주공아파트)
유시환
경기도 부천시 중동로279번길

이 발명을 지원한 국가연구개발사업

과제고유번호	1465030180
과제번호	HI19C0842020020
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	의료데이터보호・활용기술개발(R&D)
연구과제명	분산형 인공지능 연합학습 알고리즘 개발
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

의료데이터로부터 병명을 판단하도록 학습하는 의료데이터 스플릿 학습 시스템에 있어서,

상기 병명과 관련된 의료데이터를 요청하는 중앙 서버; 및

의료기관 측에 마련된 적어도 하나 이상의 클라이언트 서버로서, 상기 클라이언트 서버는 상기 중앙 서버로부터 수신한 의료데이터 요청신호에 따라 상기 병명과 관련된 원본 의료데이터를 선택하고, 선택된 원본 의료데이터의 특징맵을 추출하여 적어도 하나 이상의 훈련 의료데이터를 생성하는 상기 클라이언트 서버;를 포함하되,

상기 중앙 서버는, 상기 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하고 학습하여 의료데이터로부터 병명을 판단하는 모델을 생성하는 의료데이터 스플릿 학습 시스템.

청구항 2

제 1항에 있어서,

상기 클라이언트 서버와 상기 중앙 서버는 CNN 알고리즘을 수행하되,

상기 클라이언트 서버는 상기 훈련 의료데이터를 생성하는 1개의 Hidden Layer로 마련되는 은닉부를 포함하고,

상기 중앙 서버는 상기 적어도 하나 이상의 훈련 의료데이터를 수집하여 학습하는 제어부를 포함하되, 상기 제어부는 적어도 하나 이상의 Hidden Layer로 마련되는 의료데이터 스플릿 학습 시스템.

청구항 3

제 2항에 있어서,

상기 제어부는,

적어도 하나 이상의 클라이언트 서버로부터 특징맵을 수신하여 합산하고, 합산 값을 컨벌루션 연산하여 특징맵을 재생성하는 특징맵 연산부;

상기 특징맵 연산부가 재생성한 특징맵을 활성화 함수의 입력값으로 입력하여 연산하는 활성화 함수 연산부;

상기 활성화 함수 연산부의 연산값에 대하여 Max Pooling 연산을 이용하여 차원을 축소하는 풀링 연산부;

상기 풀링 연산부의 연산값과 미리 정해진 목표 출력값을 이용하여 손실값을 연산하는 손실값 연산부; 및

상기 손실값의 연산을 통해 각 연산부의 매개변수에 대한 수정 값을 획득하는 수정값 획득부;를 포함하는 의료데이터 스플릿 학습 시스템.

청구항 4

제 3항에 있어서,

상기 중앙 서버는,

상기 수정값 획득부로부터 획득한 매개 변수에 대한 수정값을 이용하여 상기 매개 변수를 업데이트하는 매개변수 업데이트부;를 더 포함하는 의료데이터 스플릿 학습 시스템.

청구항 5

제 3항에 있어서,

상기 활성화 함수 연산부는,

Sigmoid 함수 또는 ReLu함수를 이용하여 상기 특징벡터를 연산하는 의료데이터 스플릿 학습 시스템.

청구항 6

의료데이터로부터 병명을 판단하도록 학습하는 의료데이터 스플릿 학습 시스템의 제어방법에 있어서,

중앙 서버는 상기 병명과 관련된 의료데이터를 요청하고,

의료기관 측에 마련된 적어도 하나 이상의 클라이언트 서버는 상기 중앙 서버로부터 수신한 의료데이터 요청신호에 따라 상기 병명과 관련된 원본 의료데이터를 선택하며, 선택된 원본 의료데이터의 특징맵을 추출하여 적어도 하나 이상의 훈련 의료데이터를 생성하고,

상기 중앙 서버는 상기 적어도 하나 이상의 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하고 학습하여 상기 의료데이터로부터 병명을 판단하는 모델을 생성하는 의료데이터 스플릿 학습 시스템의 제어방법.

청구항 7

제 6항에 있어서,

상기 중앙 서버가 상기 적어도 하나 이상의 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하여 학습하는 것은,

특징맵 연산부가 적어도 하나 이상의 클라이언트 서버로부터 특징맵을 수신하여 합산하고, 합산 값을 컨벌루션 연산하여 특징맵을 재생성하고,

활성화 함수 연산부가 상기 특징맵 연산부가 재생성한 특징맵을 활성화 함수의 입력값으로 입력하여 연산하며,

풀링 연산부가 상기 활성화 함수 연산부의 연산값에 대하여 Max Pooling 연산을 이용하여 차원을 축소하고,

손실값 연산부가 상기 풀링 연산부의 연산값과 미리 정해진 목표 출력값을 이용하여 손실값을 연산하며,

수정값 획득부가 상기 손실값의 연산을 통해 각 연산부의 매개변수에 대한 수정 값을 획득하는 의료데이터 스플릿 학습 시스템의 제어방법.

청구항 8

제 7항에 있어서,

매개변수 업데이트부가 상기 수정값 획득부로부터 획득한 매개 변수에 대한 수정값을 이용하여 상기 매개 변수를 업데이트하는 것을 더 포함하는 의료데이터 스플릿 학습 시스템의 제어방법.

청구항 9

제 6항 내지 제 8항 중 어느 하나의 항에 따른 상기 데이터 스플릿 학습 시스템의 제어방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 저장 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 의료데이터 스플릿 학습 시스템과 이의 제어방법 및 그 방법을 수행하기 위한 기록매체에 관한 것으로, 보다 상세하게는 의료 데이터로부터 병명을 판단하도록 학습하는 의료데이터 스플릿 학습 시스템과 이의 제어방법 및 그 방법을 수행하기 위한 기록매체에 관한 것이다.

배 경 기 술

[0002] 딥러닝은 사물이나 데이터를 군집화하거나 분류하는데 사용하는 기술이다.

[0003] 딥러닝 기술은 학습시키는 데이터 표본의 개수가 결과의 정확성에 큰 영향을 준다. 그러므로 딥러닝을 위해서는 충분한 양의 데이터 확보와 데이터마이닝 작업이 선행되어야 한다. 데이터마이닝이란 데이터 안에서 수학모델 및 통계 기법 등을 활용하여 규칙 혹은 패턴을 찾아내는 일련의 작업으로, 데이터를 분류, 군집화하는데 사용하거나 데이터의 연관성, 연속성 등을 찾고 미래를 예측하는 것이 목적이다.

[0004] 딥러닝은 각 자료의 가중치(weight)를 이용해 확률 계산을 하며, 그 과정을 여러 번 반복하여 결과의 정확성을 높인다. 축적된 데이터는 합성곱 신경망(Convolution Neural Network : CNN)을 사용하여 학습한다.

[0005] 다만, 딥러닝 학습에 사용되는 데이터에 대하여 개인 정보 유출 문제가 발생할 수 있고, 충분한 양의 데이터 확보가 없다면 딥러닝 모델의 과적합(Overfitting)이 발생할 수 있다는 문제점이 존재한다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) KR 10-2279376

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자 하는 기술적 과제는 원본 의료데이터의 특징맵을 추출하고, 훈련 의료데이터를 생성하여 학습하는 의료데이터 스플릿 학습 시스템과 이의 제어방법 및 그 방법을 수행하기 위한 기록매체를 제공하는 것이다.

[0008] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0009] 본 발명의 일측면은, 의료데이터로부터 병명을 판단하도록 학습하는 의료데이터 스플릿 학습 시스템에 있어서, 상기 병명과 관련된 의료데이터를 요청하는 중앙 서버; 및 의료기관 측에 마련된 적어도 하나 이상의 클라이언트 서버로서, 상기 클라이언트 서버는 상기 중앙 서버로부터 수신한 의료데이터 요청신호에 따라 상기 병명과 관련된 원본 의료데이터를 선택하고, 선택된 원본 의료데이터의 특징맵을 추출하여 적어도 하나 이상의 훈련 의료데이터를 생성하는 상기 클라이언트 서버;를 포함하되, 상기 중앙 서버는, 상기 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하고 학습하여 의료데이터로부터 병명을 판단하는 모델을 생성할 수 있다.

[0010] 또한, 상기 클라이언트 서버와 상기 중앙 서버는 CNN 알고리즘을 수행하되, 상기 클라이언트 서버는 상기 훈련 의료데이터를 생성하는 1개의 Hidden Layer로 마련되는 은닉부를 포함하고, 상기 중앙 서버는 상기 적어도 하나 이상의 훈련 의료데이터를 수집하여 학습하는 제어부를 포함하되, 상기 제어부는 적어도 하나 이상의 Hidden Layer로 마련될 수 있다.

[0011] 또한, 상기 제어부는, 적어도 하나 이상의 클라이언트 서버로부터 특징맵을 수신하여 합산하고, 합산 값을 컨벌루션 연산하여 특징맵을 재생성하는 특징맵 연산부; 상기 특징맵 연산부가 재생성한 특징맵을 활성화 함수의 입력값으로 입력하여 연산하는 활성화 함수 연산부; 상기 활성화 함수 연산부의 연산값에 대하여 Max Pooling 연산을 이용하여 차원을 축소하는 풀링 연산부; 상기 풀링 연산부의 연산값과 미리 정해진 목표 출력값을 이용하여 손실값을 연산하는 손실값 연산부; 및 상기 손실값의 연산을 통해 각 연산부의 매개변수에 대한 수정 값을

획득하는 수정값 획득부;를 포함할 수 있다.

[0012] 또한, 상기 중앙 서버는, 상기 수정값 획득부로부터 획득한 매개 변수에 대한 수정값을 이용하여 상기 매개 변수에 대한 업데이트를 수행하는 매개변수 업데이트부;를 더 포함할 수 있다.

[0013] 또한, 상기 활성화 함수 연산부는, Sigmoid 함수 또는 ReLu함수를 이용하여 상기 특징벡터를 연산할 수 있다.

[0014] 본 발명의 다른 일측면은, 의료데이터로부터 병명을 판단하도록 학습하는 의료데이터 스플릿 학습 시스템의 제어방법에 있어서, 중앙 서버는 상기 병명과 관련된 의료데이터를 요청하고, 의료기관 측에 마련된 적어도 하나 이상의 클라이언트 서버는 상기 중앙 서버로부터 수신한 의료데이터 요청신호에 따라 상기 병명과 관련된 원본 의료데이터를 선택하며, 선택된 원본 의료데이터의 특징맵을 추출하여 적어도 하나 이상의 훈련 의료데이터를 생성하고, 상기 중앙 서버는 상기 적어도 하나 이상의 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하고 학습하여 상기 의료데이터로부터 병명을 판단하는 모델을 생성할 수 있다.

[0015] 또한, 상기 중앙 서버가 상기 적어도 하나 이상의 클라이언트 서버로부터 상기 적어도 하나 이상의 훈련 의료데이터를 수집하여 학습하는 것은, 특징맵 연산부가 적어도 하나 이상의 클라이언트 서버로부터 특징맵을 수신하여 합산하고, 합산 값을 컨벌루션 연산하여 특징맵을 재생성하고, 활성화 함수 연산부가 상기 특징맵 연산부가 재생성한 특징맵을 활성화 함수의 입력값으로 입력하여 연산하며, 풀링 연산부가 상기 활성화 함수 연산부의 연산값에 대하여 Max Pooling 연산을 이용하여 차원을 축소하고, 손실값 연산부가 상기 풀링 연산부의 연산값과 미리 정해진 목표 출력값을 이용하여 손실값을 연산하며, 수정값 획득부가 상기 손실값의 연산을 통해 각 연산부의 매개변수에 대한 수정 값을 획득할 수 있다.

[0016] 또한, 매개변수 업데이트부가 상기 수정값 획득부로부터 획득한 매개 변수에 대한 수정값을 이용하여 상기 매개변수에 대한 업데이트를 수행하는 것을 더 포함할 수 있다.

[0017] 본 발명의 또 다른 일측면은, 컴퓨터로 판독 가능한 저장 매체에는 의료데이터 스플릿 학습 시스템의 제어방법을 수행하기 위한 컴퓨터 프로그램이 기록될 수 있다.

발명의 효과

[0018] 상술한 본 발명의 일측면에 따르면, 복수개의 클라이언트 서버로부터 원본 의료데이터의 특징맵을 수신하여 중앙 서버가 학습하므로, 개인 정보 유출 방지 및 딥러닝 모델의 과적합(Overfitting) 문제점을 해결할 수 있다.

도면의 간단한 설명

[0019] 도1은 본 발명의 일 실시예에 따른 의료데이터 스플릿 학습 시스템을 나타내는 개념도이다.

도2는 클라이언트 서버의 제어블록도를 나타내는 개념도이다.

도3은 중앙 서버의 제어블록도를 나타내는 개념도이다.

도4는 CNN 알고리즘이 클라이언트 서버와 중앙 서버에서 분리되어 수행되는 과정을 나타내는 도면이다.

도5는 원본 데이터와 훈련 데이터를 나타내는 도면이다.

도6은 Covid-19의 확진 판별 정확도를 나타내는 표이다.

도7은 골절 판별 정확도를 나타내는 표이다.

도8 내지 도9는 손실값 연산부가 연산한 손실값을 나타낸 표이다.

도10은 본 발명의 일 실시예에 따른 의료데이터 스플릿 학습 시스템의 제어방법을 나타내는 전체적인 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있는 특정 형상, 구조 및 특성은 일 실시예와 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범

위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

[0021] 이하, 도면들을 참조하여 본 발명의 바람직한 실시예들을 보다 상세하게 설명하기로 한다.

[0022] 도1은 본 발명의 일 실시예에 따른 의료데이터 스플릿 학습 시스템을 나타내는 개념도이다.

[0023] 의료데이터 스플릿 학습 시스템(1)은 클라이언트 서버(100) 및 중앙 서버(200)를 포함할 수 있다.

[0024] 의료데이터 스플릿 학습 시스템(1)은 딥러닝 학습 알고리즘을 기초로 의료데이터로부터 병명을 판단하도록 학습하는 시스템일 수 있다.

[0025] 의료데이터는 환자의 내시경 이미지, CT 및 MRI 이미지 등을 포함할 수 있다. 병명판단은 COVID-19 확진여부, 골절여부, 콜레스테롤 수치 등을 포함할 수 있다. 딥러닝 학습 알고리즘은 심층 신경망 (Deep Neural Network, DNN), 합성곱 신경망(Convolutional Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN), 제한 볼츠만 머신 (Restricted Boltzmann Machine, RBM), 심층 신뢰 신경망 (Deep Belief Network, DBN), 심층 Q-네트워크 (Deep Q-Networks) 등의 알고리즘들이 있다. 이 때, 의료데이터 스플릿 학습 시스템(1)은 합성곱 신경망(Convolutional Neural Network, CNN)을 기초로 학습하는 시스템일 수 있다.

[0026] 중앙 서버(200)는 질병을 나타내는 MRI, CT 이미지의 의료데이터를 적어도 하나 이상의 클라이언트 서버(100)에 요청할 수 있다. 클라이언트 서버(100)는 중앙 서버(200)로부터 수신한 의료데이터 요청신호에 따라 해당 질병을 나타내는 원본 의료데이터를 선택할 수 있다. 이 때, 원본 의료데이터는 클라이언트 서버(100)에 저장된 것일 수 있다. 클라이언트 서버(100)는 원본 의료데이터로부터 특징맵을 추출하여 중앙 서버(200)에 전송할 수 있다. 이 때, 추출된 특징맵은 중앙 서버(200)의 훈련 의료데이터로 이용될 수 있는 것이다. 이 때, 복수개의 클라이언트 서버(100)간에는 개인정보 보호를 위해 원본 의료데이터를 송수신하지 않는다.

[0027] 클라이언트 서버(100)에서는 CNN 알고리즘 중에서 특징맵을 추출하는 과정만 수행될 수 있고, 중앙 서버(200)에서는 특징맵을 추출하고, 활성화 함수를 이용하여 변환할 수 있으며, 풀링 연산을 통해 차원을 축소할 수 있는 CNN 알고리즘이 수행될 수 있다.

[0028] 클라이언트 서버(100)에서 이루어지는 CNN 알고리즘은 1개의 Hidden Layer에서 수행되고, 중앙 서버(200)에서 이루어지는 CNN 알고리즘은 적어도 1개 이상의 Hidden Layer에서 수행될 수 있는 것이다.

[0029] 도2는 클라이언트 서버의 제어블록도를 나타내는 개념도이다.

[0030] 클라이언트 서버(100)는 입력부(110), 통신부(120), 제어부(130), 저장부(140) 및 출력부(150)를 포함할 수 있고, 제어부(130)는 의료데이터 선택부(131) 및 은닉부(132)를 포함할 수 있다.

[0031] 입력부(110)는 병원 측으로부터 환자 개개인의 의료데이터가 입력될 수 있다.

[0032] 통신부(120)는 중앙 서버(200)와 통신할 수 있고, 훈련 의료데이터를 중앙 서버(200)에 전송하도록 마련될 수 있다.

[0033] 저장부(140)는 입력부(110)로부터 입력된 의료데이터가 저장될 수 있다.

[0034] 출력부(150)는 원본 의료데이터로부터 추출된 특징맵을 출력할 수 있다.

[0035] 의료 데이터 선택부(131)는 중앙 서버(200)로부터 수신한 요청신호에 따라 요청받은 원본 의료데이터를 선택할 수 있다.

[0036] 은닉부(132)는 CNN 알고리즘이 수행될 수 있다. 은닉부(132)는 CNN 알고리즘이 수행될 수 있도록 Input Layer와 1개의 Hidden Layer로 마련될 수 있다. Input Layer에 CNN 알고리즘이 수행될 원본 의료데이터가 입력될 수 있다. Hidden Layer에서 입력된 원본 의료데이터에 대한 특징맵(feature map)을 추출하여 원본 의료데이터를 변형시킬 수 있다. 이를 통해, 원본 의료데이터의 보호가 가능하다. 특징맵(fc)을 추출하는 과정은 아래의 수학식 1을 통해 이루어진다.

[0037] [수학식 1]

$$net_{mn}^{l} = \sum_{i=0}^{size^{l}-1} \sum_{j=0}^{size^{l}-1} (O_{m+i,n+j}^{l-1} \cdots w_{i,j}^{l} + b^{l})$$

[0038]

[0039] $$\triangleright\ f_c = \mathrm{Conv}(O^{l-1}, w^{l}, I_n, l_a) = net_{I_n,l_a}^{l}$$

[0040] 여기에서, l은 레이어를 의미하고, size^l은 레이어의 사이즈를 의미하며, In은 input data의 수를 의미하고, Ia는 label의 수를 의미하며, O는 Output Convolution Layer를 의미하고, w는 가중치를 의미하며, b는 특징맵의 편향 bias를 의미한다.

[0041] 도3은 중앙 서버의 제어블록도를 나타내는 개념도이다.

[0042] 중앙 서버(200)는 통신부(210), 제어부(220), 저장부(230) 및 출력부(240)를 포함할 수 있고, 제어부(220)는 특징맵 연산부(221), 활성화 함수 연산부(222), 풀링 연산부(223), 손실값 연산부(224), 수정값 획득부(225) 및 매개변수 업데이트부(226)를 포함할 수 있다.

[0043] 통신부(210)는 클라이언트 서버(100)와 통신할 수 있고, 복수개의 클라이언트 서버(200)로부터 수신한 특징맵을 수신할 수 있다.

[0044] 제어부(220)는 클라이언트 서버(100)로부터 적어도 하나 이상의 훈련 의료데이터를 수집하고 학습하여 의료데이터로부터 병명을 판단하는 모델을 생성할 수 있다.

[0045] 출력부(240)는 학습한 결과로서 입력된 의료데이터로부터 판단된 병명을 출력할 수 있다.

[0046] 저장부(230)는 학습한 결과로서, 특정 병명을 가진 의료데이터와 특정 병명을 매칭하여 저장할 수 있다.

[0047] 특징맵 연산부(221)는 복수개의 클라이언트 서버(200)로부터 수신한 특징맵을 컨캐터네이팅(Concatenating)할 수 있다. 특징맵 연산부(221)는 컨캐터네이팅한 특징맵을 컨벌루션 연산을 수행하여, 특징맵을 재생성할 수 있다. 특징맵 연산부(221)가 특징맵을 재생성하는 과정은 아래의 수학식 2를 통해 이루어질 수 있다.

[0048] [수학식 2]

[0049] (1) $$\sum_{k=1}^{n} f_c^{k}$$

$$net_{mn}^{l} = \sum_{i=0}^{size^{l}-1} \sum_{j=0}^{size^{l}-1} (O_{m+i,n+j}^{l-1} \cdots w_{i,j}^{l} + b^{l})$$

[0050] (2)

[0051] $$\triangleright\ f_c = \mathrm{Conv}(O^{l-1}, w^{l}, I_n, l_a) = net_{I_n,l_a}^{l}$$

[0052] (1)은 복수개의 클라이언트 서버(100)로부터 수신한 특징맵을 컨캐터네이팅하는 과정을 나타내는 식이다.

[0053] (2)는 컨캐터네이팅한 특징맵을 이용하여 컨벌루션 연산을 수행한 결과값을 나타낸 식이다. 즉, (2)는 재생성된 특징맵을 의미할 수 있다.

[0054] 이 때, In은 input data의 수이고, Ia는 label의 수이며, O는 Output Convolution Layer이고, w는 가중치를 의미한다.

[0055] 활성화 함수 연산부(222)는 특징맵 연산부(221)가 재생성한 특징맵을 활성화 함수의 입력값으로 입력하여 연산할 수 있다. 이 때, 활성화 함수로 Sigmoid 함수 또는 ReLu함수를 이용할 수 있다.

[0056] 본 발명에서는 의료데이터가 MURA(MUsculoskeletal Radiographs), COVID-19에 관한 CT데이터인 경우에는 Sigmoid 함수를 이용하였고, Cholesterol에 관한 데이터인 경우에는 ReLu함수를 이용하였다.

[0057] 풀링 연산부(223)는 활성화 함수 연산부(222)의 연산값에 대하여 Max Pooling 연산을 이용하여 차원을 축소할

수 있다. 풀링 연산부(223)의 목적은 데이터의 차원의 크기를 축소하기 위한 것이다. 즉, 풀링 연산은 데이터에서 세로 및 가로 방향 공간의 크기를 줄이는 연산이다. 풀링 연산 방법은 다양한 형식(예: 평균, 중간 값, 최댓값, 최소값 등)이 있고, 본 발명의 실시예에서 이용하는 풀링 방법은 최댓값 풀링 연산이다. 최댓값 풀링 연산(max polling)을 이용하여 이미지의 제한 영역에서 최댓값을 추출하고, 데이터의 잡음을 제거할 수 있으며, 데이터가 줄어드는 과정에서 오버 피팅(over fitting)을 방지할 수 있다. 풀링 연산은 아래의 수학식3을 통해 이루어질 수 있다.

[0058] [수학식 3]

$$\mathrm{Pool}(x,i,j) = \frac{\sum_{m=1}^{size^l}\sum_{m=1}^{size^l} x^{l-1}_{size^l\times(i-1)+m,\,size^l\times(j-1)+n}}{size^l \times size^l};$$

[0059]

[0060] $$\triangleright\ f_p = \mathrm{Pool}(f_c, I_m, I_a)$$

[0061] 여기서, x는 풀링 레이어에서 입력되는 행렬을 의미하고, l은 레이어를 의미하며, i는 행렬에 마련된 행을 의미하고, j는 행렬에 마련된 열을 의미하며, size^l은 레이어의 사이즈를 의미한다. Im은 풀링 레이어의 Input data의 개수이며, Ia는 label의 개수를 의미할 수 있다.

[0062] 손실값 연산부(224)는 풀링 연산부(223)의 연산값과 미리 정해진 목표 출력값을 이용하여 손실값을 연산할 수 있다. 손실값 연산은 아래의 수학식 4의 MSLE, 수학식 5의 RMSLE 또는 수학식 6의 sMAPE를 이용하여 연산할 수 있다. 이 때, 미리 정해진 목표 출력값은 GT(Ground Truth)일 수 있다. 이 때, GT는 원본 의료데이터를 중앙 서버(200)의 Hidden Layer에서 컨벌루션 연산을 수행하고, 컨벌루션 연산값을 기초로 Max Pooling을 수행한 값일 수 있다.

[0063] [수학식 4]

$$MSLE = \frac{1}{N}\sum_{i=0}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2$$

[0064]

[0065] [수학식 5]

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2}.$$

[0066]

[0067] [수학식 6]

$$sMAPE = \frac{100\%}{n}\sum_{i=1}^{N}\left(\frac{|y_i-\hat{y}_i|}{|y_i|+|\hat{y}_i|}\right)$$

[0068]

[0069] y_i 는 풀링 연산부(223)의 연산값을 의미하고, $\hat{y}_i$ 는 미리 정해진 목표 출력값을 의미한다.

[0070] 수정값 획득부(225)는 손실값으로부터 각 연산부의 매개변수에 대한 수정 값을 획득할 수 있다. 이 때, 각 연산부의 매개변수는 w인 가중치를 의미할 수 있다.

[0071] 매개변수 업데이트부(226)는 수정값 획득부(225)로부터 획득한 매개 변수에 대한 수정값을 이용하여 매개 변수를 업데이트할 수 있다. 업데이트된 매개 변수를 이용하여 의료데이터 스플릿 학습 시스템(1)의 학습을 재수행

할 수 있다.

[0072] 도4는 CNN 알고리즘이 클라이언트 서버와 중앙 서버에서 분리되어 수행되는 과정을 나타내는 도면이다.

[0073] Hospital 1 부터 Hospital N은 복수개의 클라이언트 서버(100)로서 복수개의 병원측에 마련된 서버를 나타낸다. Centralized Server는 중앙 서버(200)를 의미할 수 있다. 복수개의 클라이언트 서버(100)는 Medical Data인 원본 의료데이터를 저장할 수 있다.

[0074] CNN 알고리즘은 클라이언트 서버(100)와 중앙 서버(200)에서 일련의 과정으로 수행될 수 있다. CNN 알고리즘을 구현하기 위해 Input Layer, 복수개의 Hidden Layer 및 Output Layer로 마련될 수 있다. 이 때, Input Layer와 1개의 Hidden Layer는 클라이언트 서버(100)에 마련되어 있고, 복수개의 Hidden Layer와 Output Layer는 중앙 서버(200)에 마련될 수 있다.

[0075] 클라이언트 서버(100)에 마련된 Input Layer에 원본 의료데이터가 선택되어 입력되고, Hidden Layer에서 컨벌루션 연산을 통해 원본 의료데이터의 특징맵을 추출할 수 있다.

[0076] 복수개의 클라이언트 서버(100)는 중앙 서버(200)에 추출된 특징맵을 전송하고, 중앙 서버(200)에 마련된 Hidden Layer에서 컨벌루션 연산, 활성화 함수 연산, 풀링 연산, 손실값 연산, 수정값 연산이 이루어질 수 있다.

[0077] 도5는 원본 데이터와 훈련 데이터를 나타내는 도면이다.

[0078] 도5는 COVID-19의 흉부 CT데이터를 나타내고 있다. (a)는 클라이언트 서버(100)의 저장부에 저장된 원본은 COVID-19의 흉부 CT 데이터를 의미할 수 있다. (b)는 클라이언트 서버(100)가 원본 의료데이터의 특징맵을 추출하여 생성한 변형된 COVID-19의 흉부 CT데이터인, 훈련 의료데이터를 의미할 수 있다. 따라서, 훈련 의료데이터를 생성하여 개인 정보를 보호할 수 있는 것이다.

[0079] 도6은 Covid-19의 확진 판별 정확도를 나타내는 표이다.

[0080] 실선은 본 발명의 실시예인 클라이언트 서버(100)와 중앙 서버(200)가 시공간적으로 분리되어 학습하는 의료데이터 스플릿 학습 시스템(1)에서 학습을 수행한 값을 의미할 수 있고, 점선은 클라이언트 서버(100)에서만 학습을 수행한 값을 의미할 수 있다.

[0081] 세로축은 CT데이터로부터 Covid-19 확진을 판별한 정확도(%)를 의미하고, 가로축은 학습 반복 횟수인 Epoch 수를 의미한다. 실선은 적은 Epoch 수인 경우에도 Covid-19의 확진 판별 정확도가 높은 수치를 나타내는 것을 알 수 있다.

[0082] 또한, COVID-19의 CT 데이터로서, batch size는 64를 이용하였고, input size로서는 64*64*1을 이용하였으며, 활성화 함수로서는 sigmoid 함수를 이용하였다.

[0083] 도7은 골절 판별 정확도를 나타내는 표이다.

[0084] Spatio-temporal은 본 발명의 실시예인 클라이언트 서버(100)와 중앙 서버(200)가 시공간적으로 분리되어 학습하는 의료데이터 스플릿 학습 시스템(1)에서 학습을 수행한 값을 의미할 수 있고, Single-client는 클라이언트 서버(100)에서만 학습을 수행한 값을 의미할 수 있다.

[0085] 세로축은 MURA 데이터로부터 골절여부를 판별한 정확도(%)를 의미하고, 가로축은 골절부위를 의미한다.

[0086] Spatio-temporal의 경우가 finger, elbow, forearm, hand, humerus, shoulder 및 wrist 부위의 골절여부를 판별한 정확도가 Single-client보다 골절 판별 정확도가 더 높은 수치를 나타내는 것을 알 수 있다.

[0087] 또한, MURA 데이터로서 Epoch는 50, batch size로서는 128, input size로서는 224*224*1 및 활성화 함수로서는 sigmoid 함수를 이용하였다.

[0088] 도8 내지 도9는 손실값 연산부가 연산한 손실값을 나타낸 표이다.

[0089] 복수개의 클라이언트 서버(100)로부터 수신한 콜레스테롤 수치를 기초로 목표 출력값과 실제 연산값을 이용하여 손실값을 계산한 결과를 나타내는 표이다.

[0090] 도8은 수학식 4 내지 수학식 6을 기초로 손실값 연산부(224)가 손실값을 연산하는 과정을 기초로 CDF를 측정한 표를 나타내었다. 도8을 참조하면, 실선은 본 발명의 실시예인 클라이언트 서버(100)와 중앙 서버(200)가 시공간적으로 분리되어 학습하는 의료데이터 스플릿 학습 시스템(1)에서 학습을 수행한 값을 의미할 수 있고, 점선

은 클라이언트 서버(100)에서만 학습을 수행한 값을 의미할 수 있다.

[0091] CDF(Cumulative Distribution Function)란, 누적 분포 함수로서 주어진 확률 변수가 특정 값보다 작거나 같은 확률을 나타내는 함수를 의미한다.

[0092] 그래프의 시작 부분에 가파른 기울기를 갖는다는 것은 낮은 값의 손실 분포가 크다는 것을 의미할 수 있고, 이는 목표 출력값에 가까운 실제 연산값을 나타내는 것을 의미한다. (a) 내지 (c)에서 표시하고 있는 실선이 목표 출력값에 가까운 실제 연산값을 나타내는 것을 알 수 있다.

[0093] 도9는 수학식 4 내지 수학식 6을 기초로 손실값 연산부(224)가 손실값을 연산하는 과정을 기초로 PDF를 측정한 표를 나타내었다. 도9를 참조하면, 실선은 본 발명의 실시예인 클라이언트 서버(100)와 중앙 서버(200)가 시공간적으로 분리되어 학습하는 의료데이터 스플릿 학습 시스템(1)에서 학습을 수행한 값을 의미할 수 있고, 점선은 클라이언트 서버(100)에서만 학습을 수행한 값을 의미할 수 있다.

[0094] PDF(Probability Density Function)란, 확률 밀도 함수로서 연속적인 변수에 의한 확률 분포 함수를 의미하며, 이는 특정 확률 변수 구간의 확률이 다른 구간에 비해 상대적으로 얼마나 높은지를 나타내는 함수를 의미한다.

[0095] (a) 내지 (c)에서는 본 발명의 실시예인 클라이언트 서버(100)와 중앙 서버(200)가 시공간적으로 분리되어 학습하는 의료데이터 스플릿 학습 시스템(1)으로 수행한 결과 값이 낮은 손실 값에 가까운 높은 확률을 나타내고 있는 것을 알 수 있다.

[0096] 또한, 콜레스테롤 수치 데이터로서 Epoch는 200, batch size로서 2048, input data로서 326032 및 활성화 함수로서 ReLu함수를 이용하였다.

[0097] 도10은 본 발명의 일 실시예에 따른 의료데이터 스플릿 학습 시스템의 제어방법을 나타내는 전체적인 흐름도이다.

[0098] 중앙 서버(200)는 복수개의 클라이언트 서버(100)에 특정 병명을 가진 의료데이터를 요청할 수 있다. 복수개의 클라이언트 서버(100)는 요청받은 병명을 가진 원본 의료데이터를 선택하고, 선택된 원본 의료데이터의 특징맵을 추출하며, 추출된 특징맵을 중앙 서버(200)에 전송할 수 있다.

[0099] 중앙 서버(200)는 복수개의 클라이언트 서버(100)로부터 수신한 특징맵을 합산하고, 합산한 특징맵에 대하여 컨벌루션 연산을 수행해 특징맵을 재생성할 수 있다.

[0100] 중앙 서버(200)는 재생성된 특징맵을 활성화 함수에 대입하고, 활성화 함수로부터 획득한 출력값을 이용하여 풀링 연산하며, 풀링 연산값과 미리 정해진 목표 출력값을 이용해 손실값을 연산할 수 있다. 중앙 서버(200)는 연산된 손실값을 이용해 매개변수에 대한 수정값을 획득할 수 있고, 획득된 매개변수 값으로 재학습하는 과정을 수행할 수 있다.

[0101] 이와 같은, 의료데이터 스플릿 학습 시스템의 제어방법은 애플리케이션으로 구현되거나 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나와 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

[0102] 컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자 기 테이프와 같은 자기 매체, CD-ROM, DVD와 같은 광기록 매체, 플롭티컬 디스 크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

[0103] 프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬 가지이다.

[0104] 이상에서 본 발명의 실시예들에 대하여 설명하였으나, 본 발명의 사상은 본 명세서에 제시되는 실시 예에 제한되지 아니하며, 본 발명의 사상을 이해하는 당업자는 동일한 사상의 범위 내에서, 구성요소의 부가, 변경, 삭제, 추가 등에 의해서 다른 실시 예를 용이하게 제안할 수 있을 것이나, 이 또한 본 발명의 사상범위 내에 든다고 할 것이다.

부호의 설명

[0105] 1: 의료데이터 스플릿 학습 시스템

100: 클라이언트 서버

110: 클라이언트 서버의 입력부

120: 클라이언트 서버의 통신부

130: 클라이언트 서버의 제어부

131: 의료 데이터 선택부

132: 은닉부

140: 클라이언트 서버의 저장부

150: 클라이언트 서버의 출력부

200: 중앙 서버

210: 중앙 서버의 통신부

220: 중앙 서버의 제어부

221: 특징맵 연산부

222: 활성화 함수 연산부

223: 풀링 연산부

224: 손실값 연산부

225: 수정값 획득부

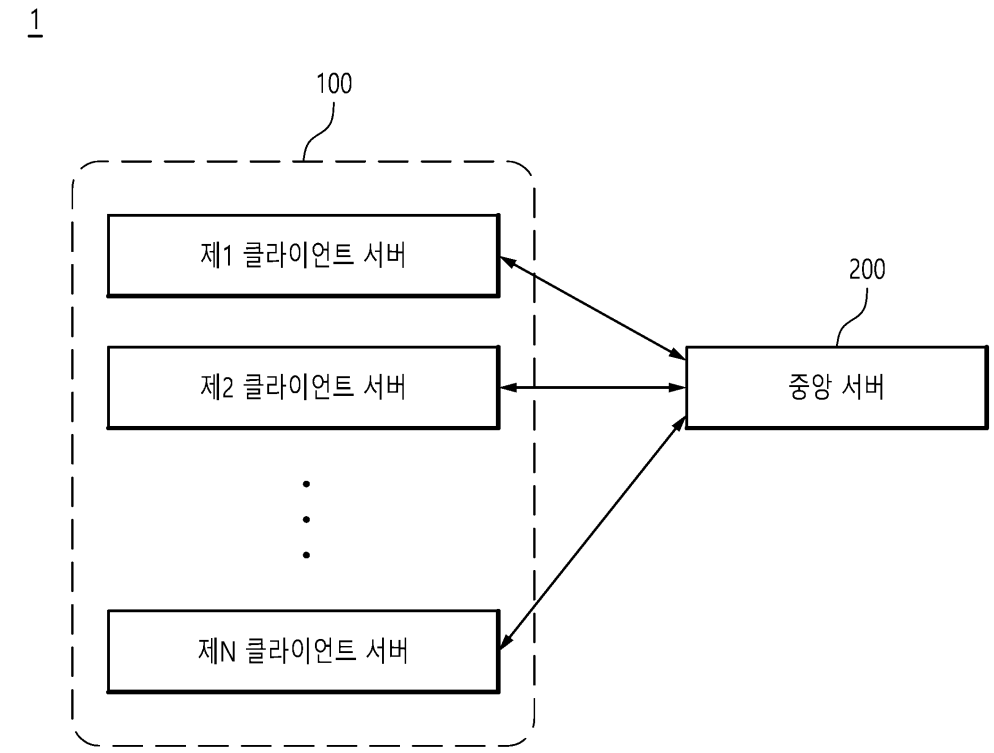
226: 매개변수 업데이트부

230: 중앙 서버의 저장부

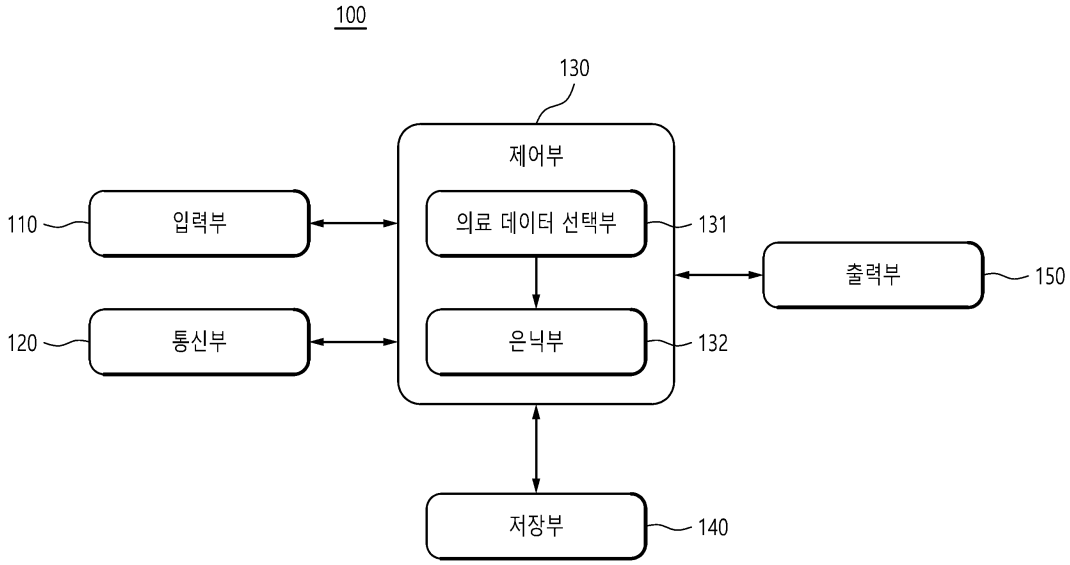
240: 중앙 서버의 출력부

도면

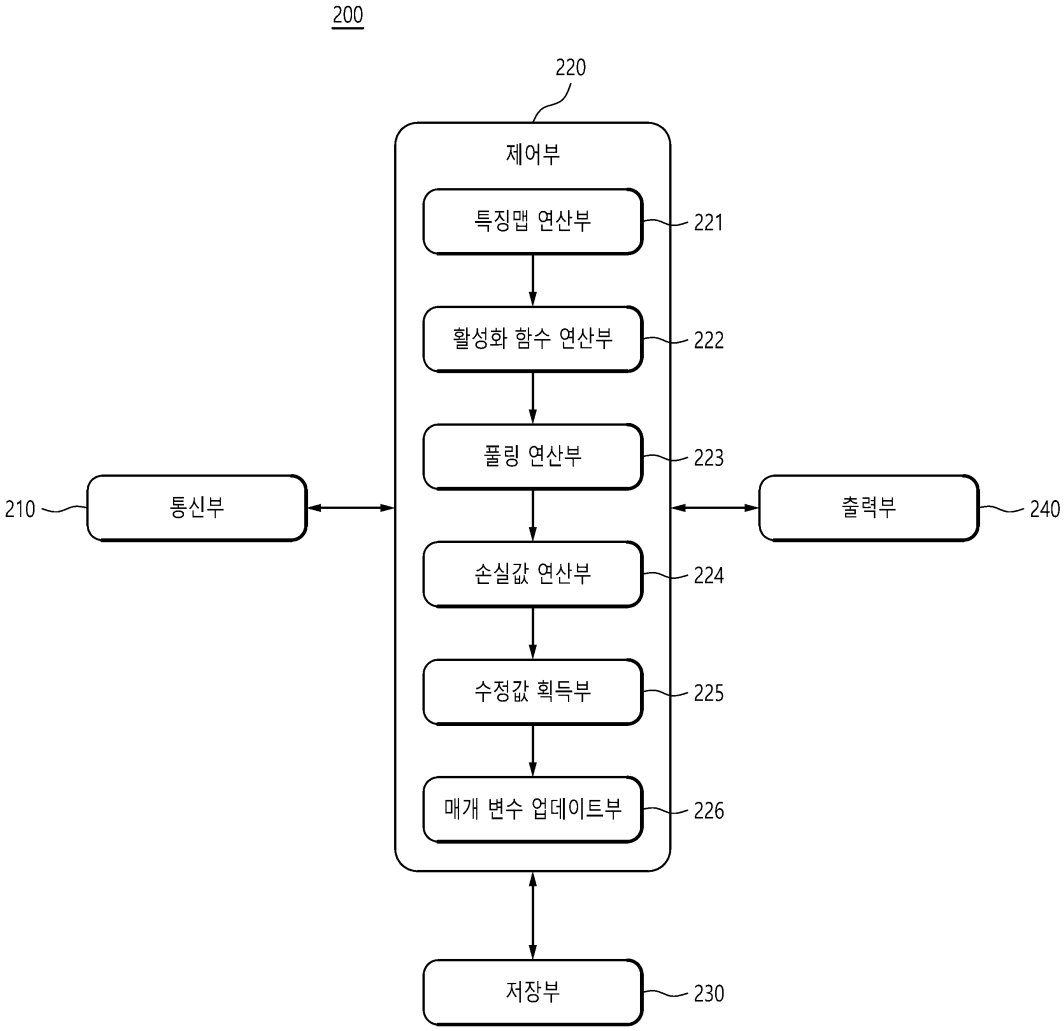
도면1


1
100
제1 클라이언트 서버
제2 클라이언트 서버
제N 클라이언트 서버
200
중앙 서버

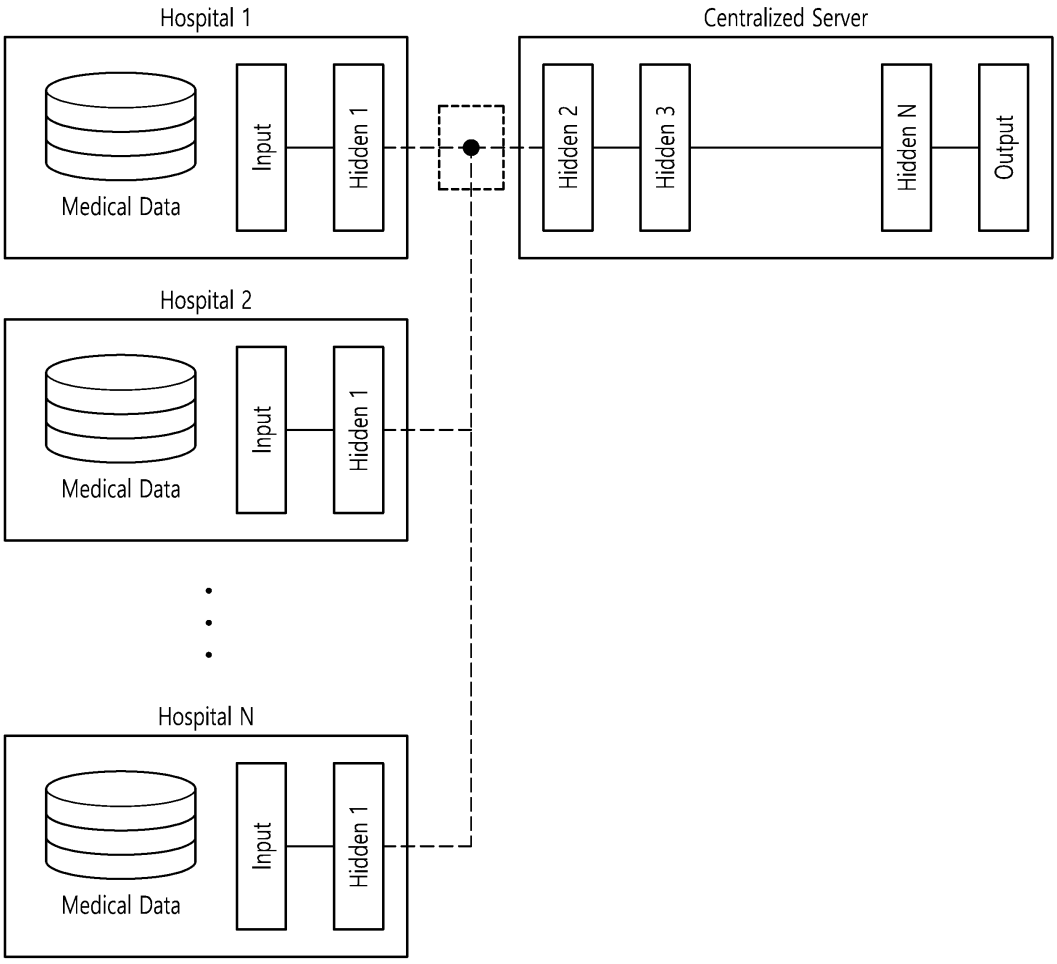
도면2


100
130
제어부
110
입력부
의료 데이터 선택부
131
출력부
150
120
통신부
은닉부
132
저장부
140

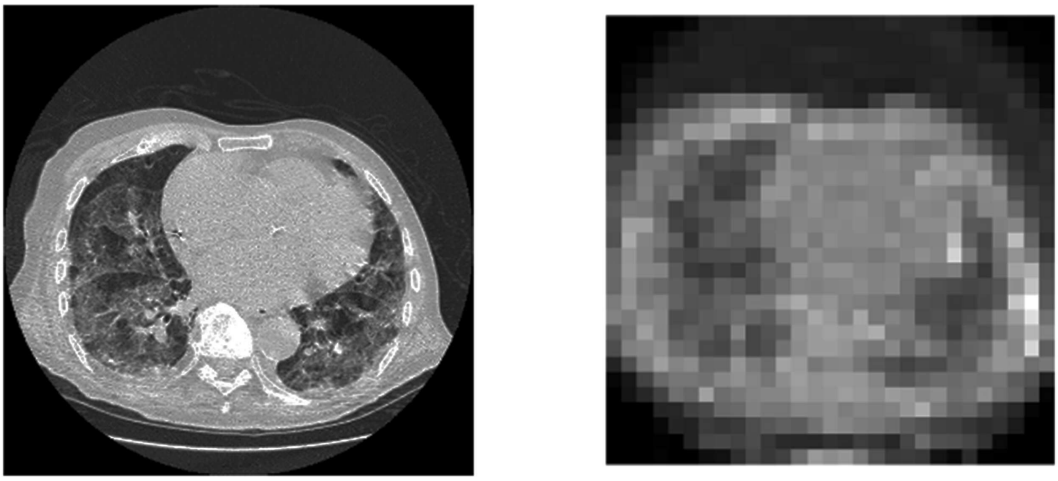
도면3


200
220
제어부
특징맵 연산부
221
활성화 함수 연산부
222
풀링 연산부
223
210
통신부
출력부
240
손실값 연산부
224
수정값 획득부
225
매개 변수 업데이트부
226
저장부
230

도면4

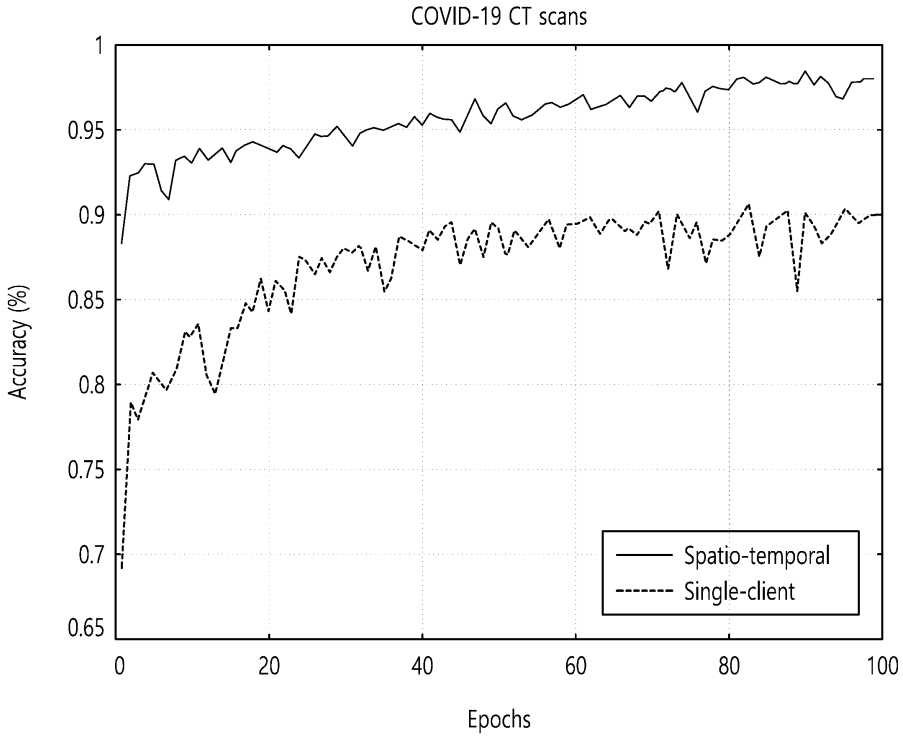

Hospital 1
Medical Data
Input
Hidden 1
Centralized Server
Hidden 2
Hidden 3
Hidden N
Output
Hospital 2
Medical Data
Input
Hidden 1
Hospital N
Medical Data
Input
Hidden 1

도면5

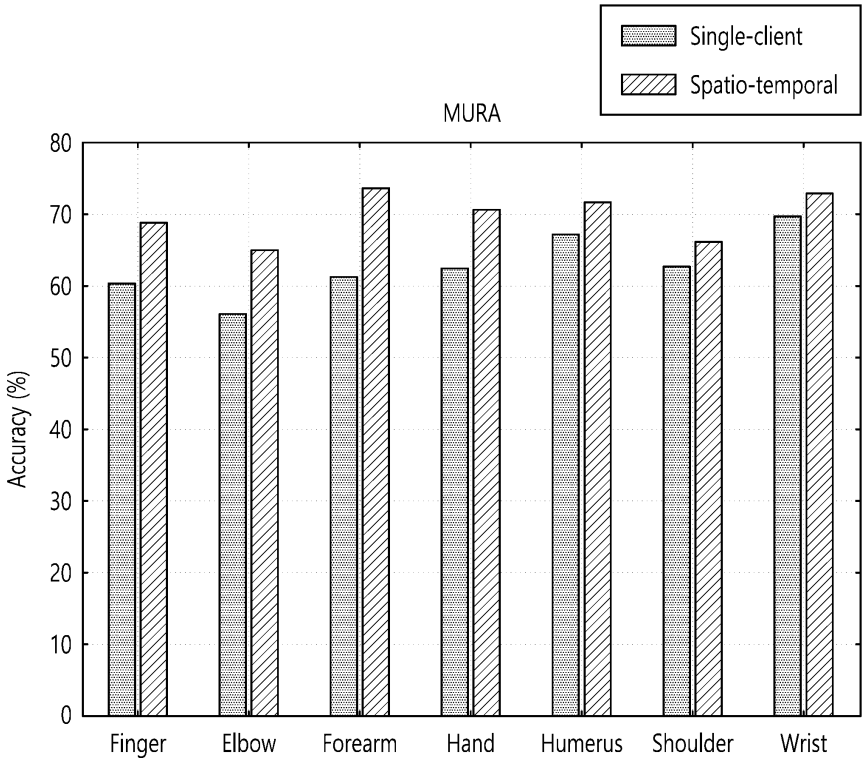


(a) (b)

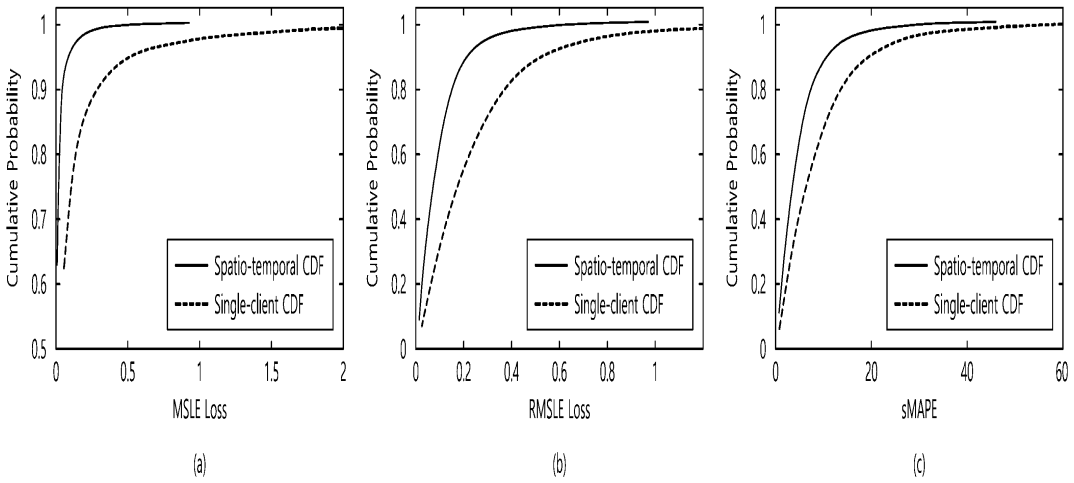
도면6


COVID-19 CT scans
Accuracy (%)
Epochs
Spatio-temporal
Single-client
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0
20
40
60
80
100

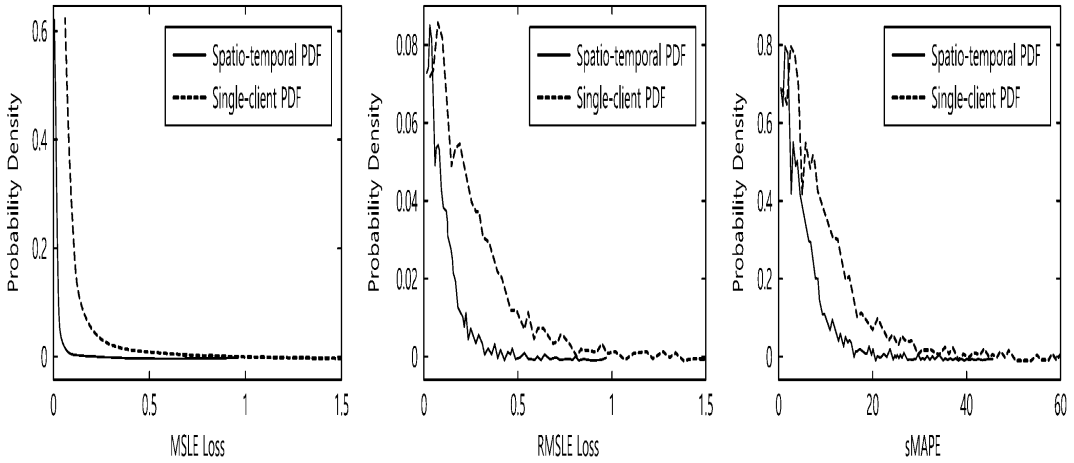
도면7


MURA
Single-client
Spatio-temporal
Accuracy (%)
80
70
60
50
40
30
20
10
0
Finger
Elbow
Forearm
Hand
Humerus
Shoulder
Wrist

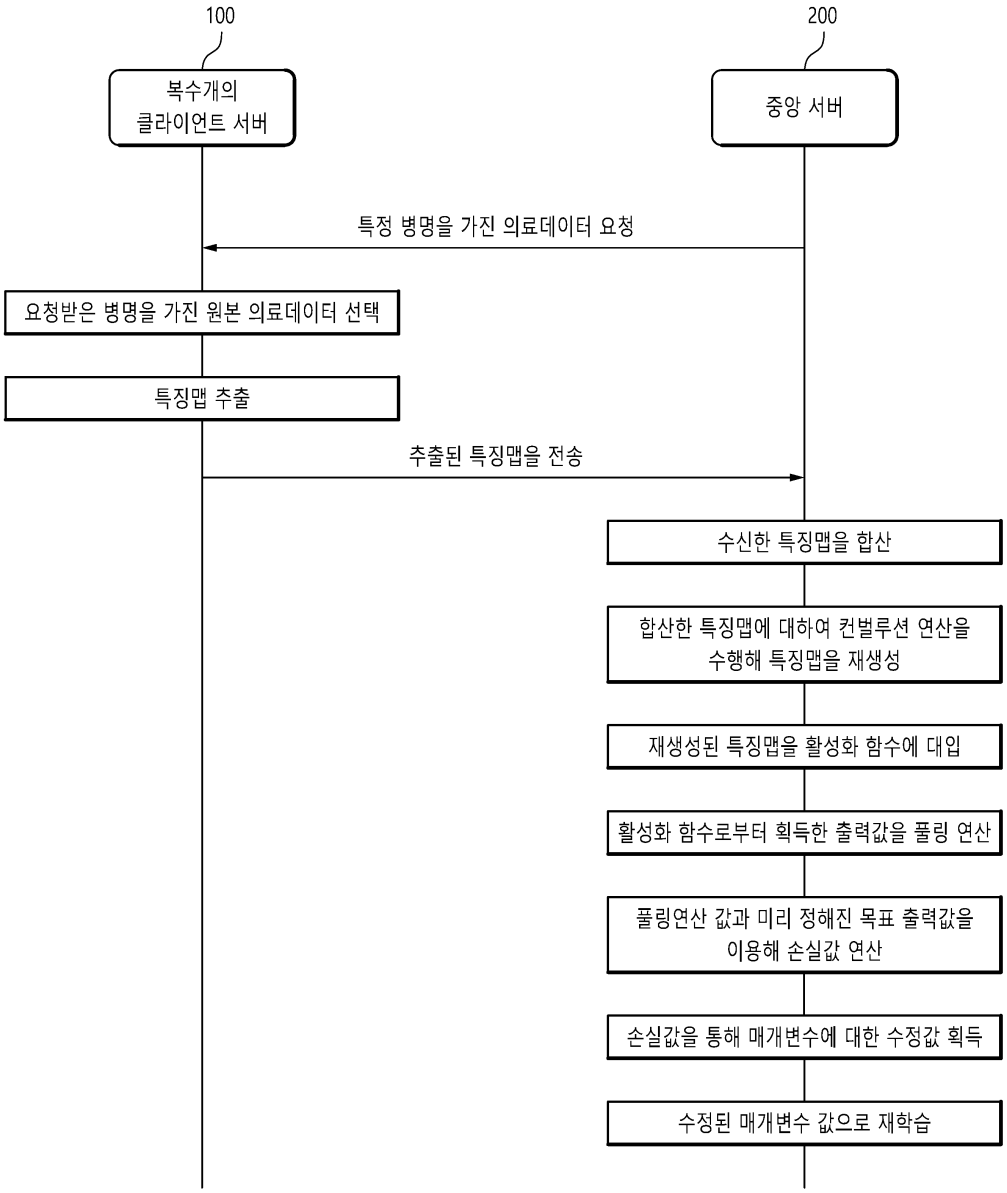
도면8


Cumulative Probability
Spatio-temporal CDF
Single-client CDF
MSLE Loss
RMSLE Loss
sMAPE
(a)
(b)
(c)

도면9


Probability Density
Spatio-temporal PDF
Single-client PDF
MSLE Loss
RMSLE Loss
sMAPE
(a)
(b)
(c)

도면10


100
200
복수개의
클라이언트 서버
중앙 서버
특정 병명을 가진 의료데이터 요청
요청받은 병명을 가진 원본 의료데이터 선택
특징맵 추출
추출된 특징맵을 전송
수신한 특징맵을 합산
합산한 특징맵에 대하여 컨벌루션 연산을
수행해 특징맵을 재생성
재생성된 특징맵을 활성화 함수에 대입
활성화 함수로부터 획득한 출력값을 풀링 연산
풀링연산 값과 미리 정해진 목표 출력값을
이용해 손실값 연산
손실값을 통해 매개변수에 대한 수정값 획득
수정된 매개변수 값으로 재학습