분산형 데이터를 학습하는 연합 학습의 의료분야 적용에 관한 연구

홍성은, 방준일, 김화종
강원대학교,
sungkenh@gmail.com, tkfka965@gmail.com, hjkim3@gmail.com

A study on the application of federated learning Distributed Data to the Medical Field

Hong Seong Eun, Bang Jun il, Kim Hwa Jong
Kangwon Univ.

요 약

최근 의료분야에서 보편적인 의료 서비스 제공을 위해 여러 기관의 진단 정보를 활용하여 질병을 판별할 수 있는 모델을 개발하기 위한 다양한 노력을 하고 있다. 하지만, 개인 정보 보호법으로 정보 주체의 정보를 함부로 사용할 수 없는 상황에서의 고성능의 판별 모델의 학습은 어려웠다. 원천 데이터를 중앙 서버로 전송하는 과정 없이 공통 모델의 공유, 로컬 학습, 가중치 공유, 공통 모델 갱신의 일련의 절차로 개인정보보호를 달성하면서 고성능의 인공지능 모델을 학습하는 신기술로 연합학습이 등장하게 되었다. 우리는 최근 대유행하고 있는 COVID-19의 이미지를 사용하여 진단할 수 있는 모델 개발에 연합 학습 방법론을 이용해보고, 분산된 환경에서 실험하고 그 성능을 평가하여, 의료 분야의 고성능 모델 학습에 연합 학습이 주는 가능성을 증명하고자 한다.

Ⅰ. 서 론

IoT와 스마트폰, 인터넷의 등장은 사람들의 생활에 전반에 스며들면서 방대한 데이터가 쌓이게 되었다. 방대한 데이터에서 의미있는 정보를 찾고자 하는 데이터 마이닝, 분석에 관한 연구, 개발이 수행되었다. 이러한 분석 기술의 발전을 위해 데이터의 품질, 표준에 관한 연구가 수행되면서 데이터의 체계적 수집 환경을 구축하였고 체계적인 방대한 데이터의 축적은 인공지능 또는 딥러닝 기술의 발전을 이룩하게 되었다.

인공지능은 인간의 삶을 편리하게 해주는 서비스에 접목되고, 인간을 뛰어넘는 결과물들을 보여주면서 많은 산업에서 주목받았고 다양한 연구들이 수행되면서 유의미한 연구 결과들이 등장하고 있다. 그와 동시에 인공지능 발전에 큰 제약 사항이 등장하였는데 바로 개인정보보호 문제였다. 데이터를 분석하여 서비스하는 데이터 비즈니스가 활성화되었던 시기에 대규모 개인정보 유출 사건이 발생하면서 EU GDPR, 개인정보 보호법이 수립되었다. [1]개인정보 보호법은 데이터를 한 곳에 수집하고 모델에 학습 입력으로 사용하는 기존 인공지능 학습 프로세스에 데이터 입력부부터 제동을 걸게 되었고, 이를 해결하고자 연합 학습 방법론이 등장하게 되었다. 연합 학습(Federated Learning)은 기존 데이터를 하나의 서버에 모아 학습하는 방식과 다르게 로컬의 데이터를 이동시키지 않고 공통 모델과 로컬 모델의 학습된 가중치를 공유하여 데이터의 이동 없이 인공지능 모델을 학습하는 방법이다. [2] 이 방법은 개인정보 민감 데이터 또는 유출 시 큰 문제가 발생할 수 있는 병원, 기업, 은행 등의 분야에서 주목받고 있다.

특히, 의료 산업에서는 과거부터 데이터 공유에 관한 관심과 의지는 있었으나, 개인 정보 유출 사건, 개인정보 보호법의 강화로 민감한 개인정보인 전자 의료 기록(EHR), 진단 기록 등은 아직도 제대로 활용되지 못하고 있다. 데이터의 공유가 불가능하여 단일 의료기관의 데이터를 사용한 인공지능 연구는 활발하게 진행되고 있지만, 인구통계학적, 지역, 전문가의 편향들이 반영되어 빅데이터를 통한 인공지능 연구들이 수행되지 못하고 있었다. 연합학습은 민감정보를 기관 간 공유 없이 대규모 데이터를 인공지능 모델로 학습할 수 있는 장점이 있다. 의료 분야에서 고성능의 대표 진단 모델은 낙후된 국가, 지역, 부족한 인력으로 인한 저품질 의료 서비스가 이루어질 수밖에 없는 환경에 대규모 데이터로 학습한 고성능의 인공지능 모델이 제공된다면 어디에서나 고품질의 의료 서비스를 받을 수 있게 된다. [3-4]

본 논문은 최근 유행하고 있는 COVID-19의 폐 스캔 데이터를 사용하여 질병을 진단하는 모델 학습에 연합 학습 기술을 사용한 실험을 수행하고, 기존의 중앙 집중식 학습 방식과의 비교 분석을 통해 연합 학습의 가능성을 증명하고자 한다.

Ⅱ. 본론

연합 학습은 기존 인공지능 학습 방법과 달리 학습할 데이터가 있는 곳(로컬 or 파티)으로 모델을 보내고 로컬 머신에서 학습된 가중치를 서버로 전송하여 서버에서는 각 로컬 머신에서 학습된 가중치를 결합하여 공통 모델을 학습하는 방식을 말한다. [5] 연합 학습은 아래 그림1과 같이 로컬 머신에서 발생하는 데이터로 하나의 공통 모델을 학습하기 위해 모델을 서버에서 공유하고 각 로컬 머신에서는 보유하고 있는 데이터로 로컬 학습을 수행한 후 학습에서 발생한 가중치를 서버로 보내 공통 모델을 갱신하는 방식으로 동작한다.

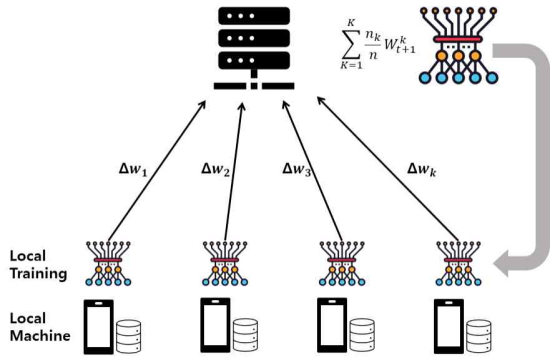

그림 1. 연합 학습 프레임워크

공통 모델을 학습하기 위해 모델을 서버에서 공유하고 각 로컬 머신에서는 보유하고 있는 데이터로 로컬 학습을 수행한 후 학습에서 발생한 가중치를 서버로 보내 공통 모델을 갱신하는 방식으로 동작한다. 이러한 과정은 그림1에 나타내고 있다.

그림2에는 연합 학습 서버(Aggregator)와 클라이언트(Party)의 기능과 구조를 나타내고 있다. Aggregator에서는 가중치를 융합하는 기능, 서버 클라이언트와의 통신을 위한 프로토콜 제어 기능, 연결 기능, 모델을 정의하여 보관하는 기능을 담당한다. Party는 데이터를 제어하고, 통신을 위한 프로토콜 핸들러, 그리고 로컬 학습을 진행하는 로컬 훈련 핸들러가 존재한다.

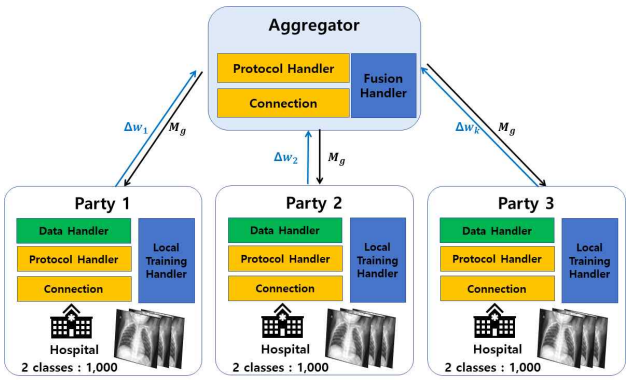

그림 2. 연합 학습 기능과 구조도

우리는 Kaggle 경진대회에서 COVID-19 Lung X-ray Scan 데이터를 사용하였으며, 각 클라이언트에 1,000개씩 데이터를 할당하였고, 연합 학습을 실험하였다. 연합 학습에서의 데이터 불균형은 불가피한 문제이며, 데이터 불균형 상태를 유지한 채로 실험을 진행하였다. 또한 각 파티는 가상의 통신 환경을 구성하여 port 번호를 부여하고 통신을 하면서 연합 학습을 수행하였다. 학습에는 CNN 기반의 MobileNetV2를 사용하였고 일부 일반화를 위한 레이어를 사용하였다. 그 결과 중앙 집중식 모델에 비해 약 -3%의 성능 차이를 보이는 학습 모델을 완성하였다. 중앙 집중식 모델은 94%의 성능을 보였고, 연합 학습으로 학습된 모델은 91%의 성능을 나타내었다. 우리는 이 결과를 통해 데이터의 공유없이도 인공지능 모델을 학습할 수 있다는 것을 보였으며, 중앙 집중식 모델과 같은 성능을 얻기 어렵지만, 좋은 성능의 모델을 만들 수 있음을 보였다.

이러한 결과들을 그림3에 나타내었다.

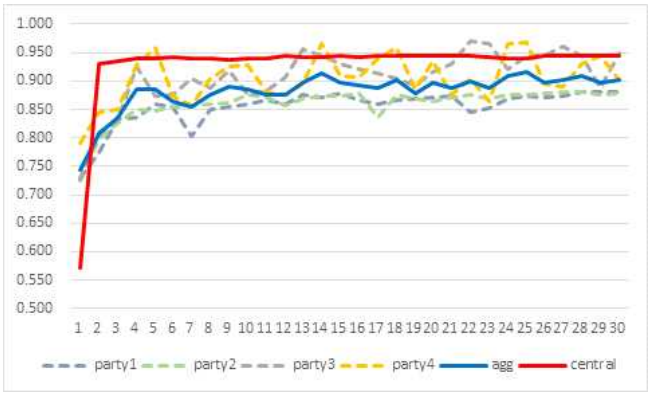

그림 3. 연합 학습과 중앙 집중식 모델의 성능 비교

Ⅲ. 결론

본 논문에서는 의료분야에서의 연합 학습 도입을 위하여 연합 학습 구조를 구현하고 이를 바탕으로 실제 데이터를 사용한 실험을 진행하였다. 특히 우리가 개발한 모델은 기존 중앙 집중식 모델에 비해 낮지만 약 2%만큼의 성능 차이를 나타내면서 데이터를 모으지 않아도 일정한 성능 달성이 가능함을 증명하였다. 그 결과 연합 학습은 개인정보보호와 인공지능이 모두 공존하는 현시대에 데이터를 공유하지 않으면서 인공지능 모델을 만들 수 있는 기술로 고려된다. 이를 방증하듯 IBM, 인텔, 엔비디아, 구글 등 다양한 기업 및 국가에서 연합 학습을 사용한 연구를 수행하고 그 결과 발표하기 시작했다.

특히 의료분야는 민감 데이터이기에 데이터 공유가 어렵고, 기관별 데이터의 편차가 매우 특별한 데이터이다. 의료분야에 연합 학습을 적용하는 연구들이 네이처지에 발표되면서 앞으로 의료분야에서의 연합 학습은 인공지능을 위한 선도 기술이 될 것으로 예상한다.

ACKNOWLEDGMENT

이 논문은 2019년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019007059)

참 고 문 헌

[1] GODDARD, Michelle, “The EU General Data Protection Regulation (GDPR): European regulation that has a global impact”, International Journal of Market Research, 2017. Vol. 59, No. 6, pp. 703-705.

[2] BONAWITZ, Keith, et al. Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046, 2019.

[3] CHEN, Yiqiang, et al. Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems, 2020, Vol. 35, No. 4, pp.83-93.

[4] DOU, Qi, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ digital medicine, 2021, Vol. 4.1, pp.1-11.