NoN-IID MNIST 데이터의 연합학습 연구

이주원[O], 방준일[*], 백종우[**], 김화종[**]
[O]강원대학교 IT대학 컴퓨터정보통신공학과,
[*]강원대학교 IT대학 컴퓨터정보통신공학과,
[**]강원대학교 IT대학 컴퓨터공학과
e-mail: {alcatraz76[O], hamster_2002[**]}@naver.com, {tkfka965[*], hjkim3[**]}@gmail.com

A Study on Federated Learning of Non-IID MNIST Data

Joowon Lee[O], Joonil Bang[*], Jongwoo Baek[**], Hwajong Kim[**]
[O]Dept. of Computer and Communications Engineering, Kangwon National University,
[*]Dept. of Computer and Communications Engineering, Kangwon National University,
[**]Dept. of Computer Science and Engineering, Kangwon National University

● 요 약 ●

본 논문에서는 불균형하게 분포된(Non-IID) 데이터를 소유하고 있는 데이터 소유자(클라이언트)들을 가정하고, 데이터 소유자들 간 원본 데이터의 직접적인 이동 없이도 딥러닝 학습이 가능하도록 연합학습을 적용하였다. 실험 환경 구성을 위하여 MNIST 손글씨 데이터 세트를 하나의 숫자만 다량 보유하도록 분할하고 각 클라이언트에게 배포하였다. 연합학습을 적용하여 손글씨 분류 모델을 학습하였을 때 정확도는 85.5%, 중앙집중식 학습모델의 정확도는 90.2%로 연합학습 모델이 중앙집중식 모델 대비 약 95% 수준의 성능을 보여 연합학습 시 성능 하락이 크지 않으며 특수한 상황에서 중앙집중식 학습을 대체할 수 있음을 보였다.

키워드: MNIST, 데이터 분포 불균형(Non-IID), 분류(Classification), 연합학습(Federated Learning)

I. Introduction

많은 분야에서 데이터 공유와 활용의 중요성이 대두되고 있다. 특히 인공지능 분야에서는 데이터의 공유를 통해 데이터양을 늘림으로써 학습하는 인공지능 모델의 성능을 향상[1]할 수 있어 양질의 데이터를 대량으로 수집하고자 노력하고 있다. 또한 인공지능 모델 개발을 원하는 데이터 소유자가 가진 데이터의 분포가 불균형할 경우(Non-IID), 인공지능 모델이 제대로 학습할 수 없어 일정한 분포를 지니도록 다양한 데이터를 수집하거나 공유하는 것이 더욱 중요하다.

다만, 원본(raw) 데이터의 공유 시 프라이버시 침해, 데이터 품질 보증과 책임 소재 문제, 지식재산권 보호 등의 문제로 인하여 데이터를 보유하고 있는 공공기관이나 기업들의 직접적인 데이터 공유는 매우 어려우며, 특히 의료, 금융 분야 등 개인정보보호 문제와 관련된 민감한 데이터를 다루는 영역에서는 직접적인 데이터 공유가 거의 불가능하다.

본 논문에서는 데이터 분포가 Non-IID한 상황에서도 데이터를 직접 공유한 것과 같은 효과를 낼 수 있도록 연합학습을 적용한 인공지능 분류 모델의 학습을 진행하였다.

이를 위하여, 독일의 Adap 社에서 발표한 연합학습 프레임워크인 Flower[2]를 사용하였고, 모델 학습을 위해 손글씨로 이루어진 숫자 이미지 데이터 세트인 MNIST 데이터 세트[3]를 사용하여 총 70,000장의 이미지를 활용하였다.

본 논문의 본론(II~III)에서는 연합학습과 MNIST 데이터의 소개 및 데이터 분할을 통한 Non-IID 환경 조성 과정에 대해 다루며, 결론부(IV)에서는 중앙집중식 학습 결과와 연합학습 결과의 비교를 통해 연합학습이 데이터 공유 및 Non-IID 상황에서의 문제점을 해결할 수 있음을 설명하였다.

II. Backgrounds

연합학습

연합학습이란 디바이스(스마트폰 등) 또는 기관(병원, 연구소 등)이 보유한 데이터(로컬 데이터)를 직접 공유하지 않으면서 인공지능 모델 학습에 간접적으로 로컬 데이터를 사용하는 일종의 분산 학습 방식이다. [4] 클라이언트는 각자 보유한 데이터로 로컬 모델을 학습시키고, 학습된 모델의 파라미터를 연합학습 서버에 전송한다. 서버는 여러 클라이언트로부터 받은 파라미터를 집계하여 글로벌 모델을

만들어 클라이언트에 배포하고, 클라이언트는 이를 파인 튜닝하는 작업(round)을 반복하여 모든 클라이언트가 성능이 개선된 글로벌 모델을 사용할 수 있다.

III. Training & Evaluation

실험 환경은 다음과 같이 구성하였다. 인공지능 모델 개발을 원하는 클라이언트 10개가 Non-IID한 데이터를 보유하고 있는 것을 가정하고, MNIST 데이터 세트를 분할하여 배포하였다. 각 클라이언트에는 약 7,000개씩의 이미지 데이터가 저장되며, 0~9까지의 숫자 중 하나를 55%, 나머지 숫자별로 5%씩을 보유하고 있다. 데이터 분포의 예시는 다음과 같다.

Table 1. 클라이언트 1의 데이터 분포 예시

Client 1	0	**1**	2	3	4	5	6	7	8	9
개수	346	**4331**	350	358	342	316	344	365	342	348
비율	5%	**55%**	5%	5%	5%	5%	5%	5%	5%	5%

클라이언트 1의 경우, 0~9까지의 숫자 중 1의 이미지를 가장 많이 보유하고 있다. 이러한 데이터 분포를 지닌 상태로 적절한 전처리 없이 인공지능 모델을 학습시키면 1에 대해 치우친 학습 결과가 도출된다.

위와 같은 클라이언트들을 10개 생성하고 연합학습을 진행하였다. 연합학습 프레임워크는 현재 배포된 프레임워크들 중 가장 사용이 간편한 Flower를 사용하였으며, 데이터를 학습할 분류 모델은 가장 기본적인 형태의 다층 퍼셉트론(MLP) 분류기를 사용하였다.

연합학습을 100라운드 진행한 결과, 분류 모델은 85.5%의 정확도를 보였다. 모든 데이터를 한곳에 모아 100 Epoch만큼 학습한 중앙집중식 학습의 경우 분류 모델은 90.2%의 정확도를 보여 연합학습 모델이 중앙집중식 모델 대비 약 95% 수준의 성능을 보이는 것을 확인하였다.

IV. Conclusions

본 논문에서는 데이터 분포가 Non-IID한 상황에서 분류 모델 학습이 가능하도록 연합학습을 적용하였다. 데이터 분포의 불균형을 인공적으로 구현하기 위하여 기존 MNIST 데이터 세트를 10개 클라이언트로 분할하여 배포하였으며, 각 클라이언트가 보유하는 데이터의 레이블별 비율은 주 레이블 55% 기타 레이블 5%로 구성하였다. 각 클라이언트가 참여하여 진행한 연합학습 모델은 85.5%의 정확도를 보였으며, 데이터를 모아 진행한 중앙집중식 학습의 정확도는 90.2%였다. 연합학습 모델의 성능은 중앙집중식 모델에 대비하여 약 95% 수준의 성능으로 연합학습을 진행하더라도 중앙집중식 학습에 비하여 성능 하락이 크지 않아 데이터 분포가 불균형하거나 데이터의 공유가 제한되는 등 특수한 상황에서 중앙집중식 학습을 대체 할 수 있음을 보였다.

다만, 본 연구에서 진행한 실험의 경우 실제 환경에 적용하기에는 아직 미흡한 면이 있다. 본 연구에서 구성한 데이터 분포의 비율은 실험자가 인위적으로 지정한 것으로, 실제 세계(Real-World)의 데이터 분포와는 다르다. 예를 들어, 클라이언트가 기타 레이블을 아예 보유하지 않고 있는 경우에는 연합학습 시 모델의 학습이 불가능하므로, 다양한 경우의 데이터 분포 상황에 관하여 연구가 필요할 것으로 보인다.

ACKNOWLEDGEMENT

본 과제(결과물)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2022RIS-005)

REFERENCES

[1] Shehzensidiq, Understanding The Importance Of Data For Machine Learning, Hackernoon, 2021.12, https://hackernoon.com/understanding-the-importance-of-data-for-machine-learning

[2] Beutel, D., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Kwing, H., Parcollet, T., Gusmão, P., & Lane, N. (2020). Flower: A Friendly Federated Learning Research Framework. arXiv preprint arXiv:2007.14390.

[3] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[4] 1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.