

Non-IID 데이터 환경에서의 적응적 연합학습 기법

이재욱*, 고한열°

Adaptive Federated Learning in Non-IID Data Environment

Jae-wook Lee*, Haneul Ko°

요약

본 논문에서는 Non-IID 데이터 환경에서 요구하는 학습 정확도를 보장하면서 낮은 수렴 시간을 제공할 수 있는 적응적 연합학습 기법인 FedA 기법을 제안하였다. 제안하는 기법의 경우 Non-IID 정도에 따라 적절한 기존의 연합학습 기법을 선택한다. 실험을 통해 FedA에서의 선택 기준을 도출하였고, 실험을 통해 도출된 선택 기준을 적용하여 FedA 기법의 우수성을 실험적으로 입증하였다.

키워드 : 연합학습, non-IID 문제, 학습 정확도
Key Words : Federated Learning, non-IID problem, training accuracy

ABSTRACT

In this paper, we propose the adaptive federated learning called FedA under non-IID data environment to guarantee the training accuracy and reduce training time. FedA adaptively selects the proper traditional federated learning scheme according to the non-IID degree. Also, we conduct the simulation to obtain the policy representing which federated learning scheme is configured according to the non-IID degree and to confirm the outperformance of our proposed scheme.

1. 서론

인공 지능 서비스를 구현하기 위해서는 타겟 서비스에 맞는 심층 모델을 학습데이터를 통해 학습하는 과정이 필수적이고, 최근 사용자의 프라이버시 침해를 보장하면서 서버에 집중되는 과도한 컴퓨팅 부하를 단말들에게 분산시킬 수 있는 연합학습 기법 (federated learning, FL)이 각광받고 있다^[1]. 연합학습 기법의 경우 단말들이 학습에 참여하기 때문에 단말들의 특성 (단말들의 학습 데이터 분포 및 컴퓨팅/통신 능력)에 따라 학습 성능 (심층 모델의 정확도 및 학습 완료 시간)이 결정되는 특징을 갖는다^[2]. 특히, 단말들이 가지고 있는 학습 데이터 분포의 특징은 학습 정확도에 큰 영향을 준다. 기령, 학습에 참여하는 단말들이 이상적인 학습 데이터 분포와 상이한 데이터 분포의 데이터들을 갖을 수록 학습 정확도가 크게 저하되는 문제가 발생한다^[2,3]. 이러한 문제를 non-independent and identically distributed 문제 (non-IID problem)라 정의하고, 해당 문제를 해결하기 위한 많은 연구들이 이루어져왔다^[2,3].

대다수의 연구들^[2]에서는 학습에 참여하는 단말들의 전체 데이터 분포가 이상적인 분포에 유사할 수 있도록 참여 단말을 선택하는 기법을 통해 non-IID 문제를 해결하였다. 하지만, 해당 방법은 각 단말들이 갖은 데이터들이 어느정도 이상적인 학습 데이터 분포와 유사한 경우에는 (non-IID 성이 낮은 경우) 학습 성능 저하 문제를 완화시킬 수 있지만, non-IID 성이 높은 경우에는 학습 저하 문제를 해결하지 못하는 문제점이 존재한다^[3]. 해당 문제를 해결하고자 [3]에서는 단말들이 순차적으로 학습함으로써 non-IID 환경에서 학습 성능 저하 문제를 해결하는 기법을 제안하였으나 순차적인 학습방법에 따라 학습시간이 증가하는 문제가 발생한다. 따라서, 본 연구에서는 non-IID 정도에 따라 단말이 동시에 학습을 수행하는 연합학습기법 (FL) 과 단말이 순차적으로 학습을 수행하는 연합학습기법 (FedS)을 선택하는 적응적 연합학습 기법을 제안한다. 또한, 시뮬레이션을 통해 제안한 연합학습의 성능을 검증한다.

* “이 논문은 2023학년도 부경대학교의 지원을 받아 수행된 연구임(202303830001).”

• First Author : Pukyong National University, Department of Information and Communication Engineering, jlee0315@pknu.ac.kr, 정희원

° Corresponding Author : Kyung Hee University, Department of Electronic Engineering, heko@khu.ac.kr, 종신희원

논문번호 : 202405-090-C-LU, Received May 6, 2024; Revised May 8, 2024; Accepted May 8, 2024

II. 적응적 연합학습 기법

본 장에서는 non-IID 환경에 따라 적응적으로 단말이 동시에 학습을 수행하는 연합학습 기법 FL 기법과 단말이 순차적으로 학습을 수행하는 FedS 기법 중 하나의 기법을 선택하여 학습을 진행하는 적응적 연합학습 알고리즘 (FedA)을 제안한다. 그림 1 (a), (b)는 FL 기법과 FedS 기법의 동작과정으로 나타낸다. 제안하는 FedA의 동작과정은 다음과 같다. 우선, 에지 서버는 단말들에게 데이터 분포 정보를 요청하고, 해당 요청에 따라 단말들은 자신이 가지고 있는 데이터의 분포를 에지 서버에 알려준다. 에지 서버는 해당 정보를 수집하여 각 단말들의 평균 non-IID 정도 (ω 값)를 수식 (1)과 같이 계산한다.

$$\omega = \frac{\sum_i KL(P_T|P_i)}{I} \quad (1)$$

수식 (1)에서 $KL(A|B)$ 함수는 분포 A와 B간의 KL divergence 값을 의미하며 P_T 와 P_i 는 학습에 이상적인 데이터 분포와 단말 i 의 데이터 분포를 의미한다. 에지 서버는 ω 값과 연합학습 선택 기준 값 θ 을 비교하여 단말들과 수행할 연합학습 기법을 선택한다. 만일 ω 값이 θ 값보다 작다면 단말들의 non-IID 상이 높지 않다는 것을 의미하기 때문에 에지 서버는 FL 기법을 선택하고, 반대의 경우 긴 학습 시간의 감수하더라도 단말들의 non-IID 데이터에 의한 학습 저하를 방지하기 위해서 FedS 기법을 선택한다. 학습 기법을 선택한 후에 단말에게 학습 방법을 설정하여 선택된 학습 기법을 수행한다.

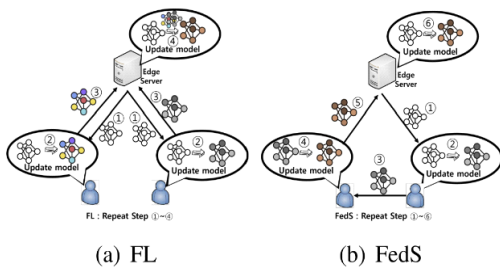


그림 1. FL 기법과 FedS 기법 동작과정
Fig. 1. Procedure of FL and FedS

III. 실험 결과

본 장에서는 우선 연합학습 선택 기준 값인 θ 을 설정

하기 위해 non-IID 환경에 따른 연합학습 기법들의 성능을 실험적으로 분석하고, 제안 기법의 성능을 분석한다. 본 실험에서는 VGG 모델을 Cifar-10 데이터^[2]를 통해 학습하고 평가하였다. 또한 단말은 10개의 단말이 연합학습에 참여하다고 가정하였고, 각 단말은 임의의 Cifar-10 데이터를 할당 받아 연합학습을 수행한다. 이때, 단말에서 학습을 위한 전송 시간과 학습 시간은 비교 계산의 편의성을 위해 각각 1msec으로 설정하였다. 끝으로 VGG 모델 기반의 분류기는 학습을 통해 학습에 참여하지 않은 데이터에 대해서 60% 이상의 분류 정확도를 요구한다고 설정하였다.

3.1 non-IID 환경에 따른 연합학습 기법 성능

그림 2의 (a)와 (b)는 ω 값에 따라 FL 기법과 FedS 기법의 학습 그래프를 나타낸다. ω 값이 클수록 단말들의 non-IID 정도가 높은 환경을 나타내며, ω 값이 0인 경우 모든 단말이 이상적인 데이터 분포를 갖는 환경이다 (IID 환경). 그림 2 (b)를 통해 단말의 non-IID 정도가 높을수록 학습을 통한 성취 정확도가 낮아지는 것을 확인할 수 있다. 특히 non-IID 정도가 심한 경우 ($\omega = 9$, 각 단말의 열개의 종류 데이터 중 두 종류의 데이터 들만 갖는 경우), 성취 정확도가 매우 낮아짐을 확인할 수 있다. 해당 결과를 통해 학습에 사용되는 전체 데이터의 분포가 IID 하더라도, 각 단말의 이질적인 데이터 분포로 인해 non-IID 문제가 심해지는 것을 확인할 수 있다. 이와 반면에 FedS는 non-IID 정도가 심한 환경에서도 성취 정확도 저하를 효율적으로 방지하는 것을 그림 2 (a)를 확인할 수 있다. 하지만, FedS와 FL의 수렴시간을 비교해보면 FL 기법이 압도적으로 짧은 것을 확인할 수 있다. 본 실험 결과를 통해 FedS와 FL의 장단점을 확인할 수 있으며, 요구하는 정확도를 충족하면서 짧은 학습 시간을 보장하기 위한 FedS와 FL 기법 선택 값 θ 을 찾을 수 있다.

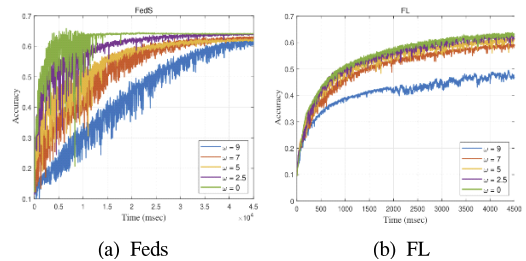


그림 2. FedS와 FL 정확도 그래프
Fig. 2. Accuracy Curves of FedS and FL

3.2 적응적 연합학습 기법 성능

그림 3 (a)와 (b)는 ω 값에 따른 각 연합학습 기법의 성취 정확도와 수렴 시간을 각각 나타낸다. 요구하는 학습 정확도는 0.6 이기 때문에 그림 2의 결과를 참고하여 θ 를 7로 설정하였다. 그림 3 (a), (b)의 결과를 보면 FedS는 항상 높은 성취 정확도를 보여주지만, 매우 긴 지연시간을 발생시킨다. 이는 FedS의 동작과정 특성상 심층 모델을 단말들이 순차적으로 학습함으로써, 모든 데이터를 모은 후에 일정 배치 양만큼 데이터를 순차적으로 학습시키는 중앙 집중적인 학습 기법과 유사한 효과를 나타내기 때문에 (즉, IID 데이터를 하나의 서버가 학습하는 효과) non-IID 문제를 해결할 수 있지만, 한시점의 하나의 배치 양 (즉, 단말이 가지고 있는 데이터 양)만큼의 데이터들만 학습에 참여하고, 다른 단말들에게 파라미터를 전송하는 지연시간 등에 의해 긴 지연시간이 필연적이다. 반대로, FL의 경우 각자의 단말의 non-IID 데이터로 동시다발적으로 학습하기 때문에 non-IID 정도에 따른 학습 성취 정확도 감소는 필연적이나, 병렬적으로 학습을 하는 특성에 의해 짧은 학습 수렴시간을 나타낸다. 이와 반대로, 제안하는 적응적 연합학습 기법인 FedA는 non-IID 정도가 높은 경우, 요구하는 학습 정확도 달성을 우선시하여 FedS 기법을 선택하고, 반대의 경우 학습속도를 높이기 위해 FL을 선택함으로써, 항상 요구하는 정확도를 보장하면서 낮은 수렴시간을 보여준다.

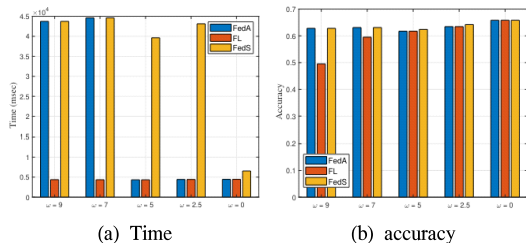


그림 3. ω 에 따른 실험 결과
Fig. 3. Effect of ω

IV. 결론

본 논문에서는 non-IID 환경에서도 요구하는 정확도를 보장하면서 낮은 수렴시간을 제공하는 적응적 연합학습 기법을 제안하였으며, 환경에 알맞은 연합학습 기법을 선택하기 위한 기준을 실험적으로 도출하고, 제안한 기술의 우수성을 실험적으로 입증하였다. 향후 연구로는 다양한 환경에서 선택 기준을 정하는 방법과 병렬적으로 FedS 기법을 수행하여 학습 속도를 줄이는 연구를 진행할 예정이다.

References

- [1] Y. Cho, et al., "A Study on the Performance Improvement of Federated Learning systems," in *Proc. KICS in Proc. KICS Winter Conf.*, pp. 837-838, Jan. 2020.
- [2] J. Lee, et al., "Data distribution-aware online client selection algorithm for federated learning in heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 1127-1136, Jan. 2023. (<https://doi.org/10.1109/TVT.2022.3205307>)
- [3] M. Duan, et al., "Self-balancing federated learning with global imbalanced data in mobile system," *IEEE Trans. Parallel and Distrib. Syst.*, vol. 32, no. 1, pp. 59-71, Jan. 2021. (<https://doi.org/10.1109/TPDS.2020.3009406>)