Non-IID 기반 연합학습 환경에서 데이터 리치 클라이언트의 영향 분석

팜칸콴(*), 김태홍(*)
(*) 충북대학교 정보통신공학부, {khanhquan, taehongkim}@cbnu.ac.kr

The Impact of Data-Rich Clients in Non-IID Federated Learning

Pham Khanh Quan(*), Taehong Kim(*)
(*) School of Information and Communication Engineering, Chungbuk National University

요 약

This paper introduces a novel method for generating a data-rich client in federated learning (FL) when dealing with non-IID data of a single class. A data-rich client refers to a client that contains a remarkable amount of training data and a diverse range of categories compared to other clients participating in the training process. The approach tackles the challenge of 1-class non-IID data by employing techniques like data augmentation, leveraging publicly available datasets, and federated transfer learning. Experimental results showed the effectiveness of the proposed approach in enhancing global and local accuracy, thus contributing to the improvement of federated learning systems in handling imbalanced data distributions.

1. Introduction

In the field of data analysis and machine learning, the assumption of Independent Identically Distributed (IID) data has been a standard practice, allowing for simpler analysis and modeling. However, in many real-world scenarios, data does not expect the IID assumption, presenting challenges and complexities in the learning process. One such case is non-IID data [1] with a single class, where each client has a unique characteristics training sample and significant imbalance among the distribution. This paper presents a novel approach to create a data-rich client in federated learning (FL) when faced with non-IID (Independent Identically Distributed) data consisting of a single class. We suggest generating a data-rich client by using some techniques such as data augmentation, leveraging publicly available datasets, and federated transfer learning. During our exploration of these methods, we prioritize the preservation of data privacy [2]. Experimental results demonstrate the effectiveness of our approach in improving global and local accuracy. Our findings contribute to enhancing the performance of federated learning systems with imbalanced data distributions.

2. Problem definition and methodologies

2.1 Single class non-IID data

In FL, single class non-IID data refers to a scenario where each client possesses data from a single class. This non-uniformity in data distribution poses challenges for training models that can effectively generalize across classes and accurately classify data from different clients. Addressing the single class non-IID data scenario requires techniques to handle imbalanced data distribution and ensure the model's convergence is not negatively affected by the different class proportions among clients.

2.2 Data augmentation

Data augmentation refers to a set of techniques used to artificially expand a dataset by applying various transformations or modifications to existing data samples. These transformations can include rotations, translations, scaling, cropping, flipping, adding noise, or changing color intensities. The purpose of data augmentation in single class non-IID data is to help the single class clients to increase the diversity of the data without collecting additional data from other clients. This also mitigates overfitting and enhances the model's robustness and performance.

2.3 Publicly available datasets

A publicly available dataset refers to a dataset that is openly accessible and can be freely used by researchers, developers, and the general public for various purposes, such as training machine learning models or conducting data analysis. If there are publicly available datasets that contain data from multiple classes and have an acceptable similarity with the single class client's datasets, we can consider using those datasets to create a data-rich client. However, we need to ensure that the publicly available datasets are compatible with our tasks and adhere to the necessary privacy and usage agreements.

3. Experimental setup

This paper aims to evaluate the impact of a data-rich client on the performance of federated learning in a non-IID environment. The data-rich client is characterized as having a substantial amount of data achieved through techniques such as data augmentation and the utilization of publicly available datasets.

3.1. Without the data-rich client

In this experimental setup, we focus on the Fashion-MNIST dataset and simulate how the non-IID scenario effect to model's performance without a data-rich client. The images represent various fashion items, such as T-shirts, dresses, shoes, bags, etc. We have a total of 10 clients in our testbed, with each client representing one class from 0 to 9. This means that each client has a subset of the dataset containing only 6000 samples from a single class.

3.2. With the data-rich client

In this experimental setup, we aim to examine the impact of the non-IID scenario on model performance while incorporating a data-rich client. Within our testbed, we have a total of 10 clients. Nine of ten clients represent individual classes from 1 to 9, meaning each client exclusively possesses a subset of the dataset containing samples from a single class. The remaining client, known as the data-rich client, stands out as it contains a comprehensive collection of 60,000 samples, including 6,000 samples from each of the 10 classes.

4. Performance analysis and results

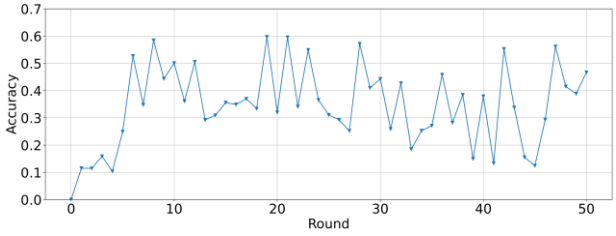

Figure 1: Global accuracy in single class non-IID without the data-rich client.

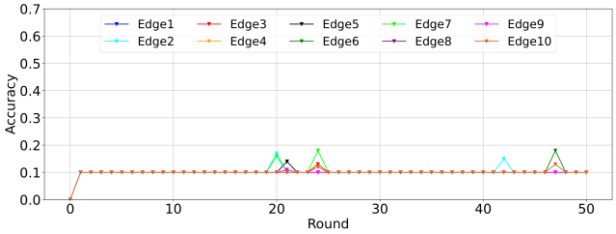

Figure 2: Local accuracy in single class non-IID without the data-rich client.

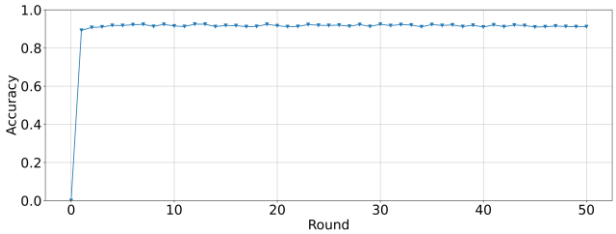

Figure 3: Global accuracy in single class non-IID with the data-rich client.

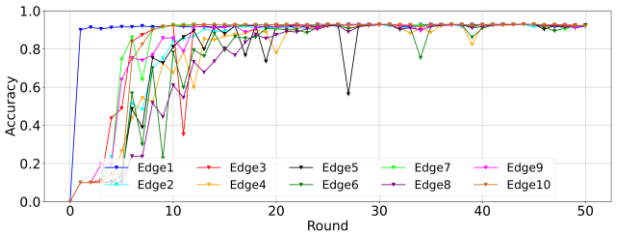

Figure 4: Local accuracy in single class non-IID with the data-rich client.

From Figure 1, it is evident that without a data-rich client, global accuracy remains limited, reaching a maximum of only 0.6, and lacks stability. Additionally, Figure 2 shows that the local accuracy for the single class non-IID scenario without a data-rich client range from 0.1 to 0.2. However, the results presented in Figures 3 and 4 demonstrate a significant improvement in both global and local accuracy when a data-rich client is introduced. After approximately 2 rounds for global accuracy and 20 rounds for local accuracy, the model accuracy can reach up to 0.9. These findings highlight the significant impact of using a data-rich client, outperforming scenarios without one. The lessons learned from these results emphasize the importance of including a data-rich client to enhance both global and local accuracy in non-IID federated learning.

5. Conclusion and limitations

In this paper, a novel approach is presented to address the impact of 1-class non-IID data by generating a data-rich client through techniques such as data augmentation, leveraging publicly available datasets, and utilizing federated transfer learning. The experimental results highlight the effectiveness of the approach in improving both global and local accuracy. These findings contribute significantly to enhancing the performance of federated learning systems dealing with non-IID data distributions. However, it should be noted that creating a data-rich client in practice can be challenging due to the lack of publicly available training samples and the need to carefully consider the quality and privacy of augmentation data.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2022R1I1A3072355).

References

[1] Zhu H, Xu J, Liu S, and Jin Y, "Federated learning on non-IID data: A survey," in Neurocomputing, vol. 465, pp.371-390, 2021.

[2] Li and Qinbin, "A survey on federated learning systems: vision, hype and reality for data privacy and protection." IEEE Transactions on Knowledge and Data Engineering, 2021.