Analyzing Federated Learning Models with IID and Non-IID Data Variations

No Author Given

Abstract. Federated Learning has emerged as a transformative approach for collaborative machine learning while preserving data privacy in decentralized settings. This paper investigates the performance of Federated Learning models in the context of the MNIST dataset, under two distinct data distribution scenarios: Independent and Identically Distributed (IID) and Non-Independent and Non-Identically Distributed (Non-IID) settings. The study begins with an overview of Federated Learning methodologies and the significance of preserving data privacy in distributed environments. It then delves into the challenges posed by data heterogeneity and non-uniform distribution across edge devices in a federated learning framework using the MNIST dataset as a case study. Experimental results demonstrate an accuracy of 87.69% in IID data settings, showcasing the model's performance under homogenous data. Additionally, in Non-IID scenarios, an accuracy of 91.6% reflects the model's adaptability to diverse data distributions across decentralized devices. Insights gained provide valuable implications for Federated Learning approaches concerning data heterogeneity, emphasizing the need for strategies to enhance model adaptability in real-world applications.

Keywords: federated learning · iid data · non-iid data

1 Introduction

Federated Learning has emerged as a groundbreaking paradigm in machine learning, enabling collaborative model training across decentralized devices while preserving data privacy. In the landscape of decentralized learning, where data is distributed across numerous edge devices, Federated Learning offers a compelling solution by allowing model training directly on user devices without the need to centralize data.

The growing prominence of Federated Learning stems from its application in diverse sectors, including healthcare, finance, and Internet of Things (IoT), where data privacy is paramount. However, Federated Learning encounters challenges when dealing with data heterogeneity and non-uniform distributions across decentralized nodes. Understanding the implications of data distribution patterns on model performance is crucial for realizing the full potential of Federated Learning in real-world scenarios.

In two different data distribution scenarios—Independent and Identically Distributed (IID) and Non-Independent and Non-Identically Distributed (Non-IID)—this

research examines the performance of federated learning models using the well-known MNIST dataset. The MNIST dataset, which consists of handwritten digit pictures, provides an example case study to investigate how heterogeneity in data distribution affects federated learning.

Our research aims to evaluate and compare the accuracy achieved by Federated Learning models in both IID and Non-IID data settings. Specifically, we present experimental results showcasing an accuracy of 87.69% in the IID data distribution scenario, highlighting the model's proficiency in handling homogeneously distributed data. Furthermore, in Non-IID data settings, our model achieves an accuracy of 91.6%, emphasizing its robustness in adapting to varying data distributions across decentralized devices.

The insights derived from this study provide valuable implications for Federated Learning methodologies in handling data heterogeneity, addressing challenges, and identifying opportunities to improve model adaptability in real-world applications. Understanding the nuances of data distributions in Federated Learning environments is crucial for advancing the efficacy and scalability of decentralized learning systems.

This introduction sets the stage by highlighting the significance of Federated Learning, the challenges posed by data heterogeneity, and the motivation behind the investigation using the MNIST dataset as a case study. It outlines the objectives and previews the experimental results, emphasizing the importance of understanding data distribution patterns in Federated Learning research.

2 Literature Review

Federated Learning (FL) has garnered significant attention due to its potential to enable collaborative model training across distributed devices while respecting data privacy. However, the conventional FL methodologies face challenges that impede their widespread adoption. Several studies have highlighted these limitations, emphasizing the need for novel approaches to enhance FL efficiency.

Existing research in the FL domain has extensively focused on communication efficiency, convergence speed, and optimization robustness. Traditional FL methods often suffer from high communication overhead due to frequent model parameter exchanges between the server and participating devices. Additionally, slow convergence rates have been observed, primarily attributed to the heterogeneity of local datasets and varying computation capabilities across devices. Moreover, ensuring robustness in optimization while preserving data privacy remains a critical concern.

Recent literature has proposed various enhancements to FL techniques to address these challenges. Studies have explored federated optimization algorithms, such as Federated Averaging and Federated Momentum, to improve convergence speed and communication efficiency. Other approaches, including differential privacy and secure aggregation protocols, aim to ensure data privacy during the collaborative training process. Despite these efforts, achieving a balance between

communication efficiency, convergence speed, and privacy preservation remains a complex challenge.

FedAvg is an iterative model averaging technique that is both conventional and practical for training deep networks in federated learning (McMahan et al., 2017) [1]. In order to solve client-drift in local updates, SCAFFOLD presents an algorithmic method that makes use of control variates (Karimireddy et al., 2020) [2]. FedLin is an algorithmic framework that addresses the issues of fuzzy communication in FL, system heterogeneity, and objective heterogeneity (Mitra et al., 2021) [3]. Using momentum-assisted stochastic gradient directions for both client and server updates, STEM functions as a stochastic two-sided momentum algorithm (Mitra et al., 2021) [3]. To address class imbalance in FL, CLIMB provides an agnostic constrained learning formulation that requires no further knowledge beyond the normal FL objective. To address distributed machine learning difficulties, MFL uses momentum gradient descent in local FL updates (Liu et al., 2020) [4]. In contrast to FedAvg's usage of stochastic gradient descent (SGD), FedOpt offers a flexible algorithmic framework that lets clients and servers choose from a variety of optimization techniques (Reddi et al) [5]. MimeLite mimics the behavior of centralized techniques on i.i.d. data by combining control variates and server-level optimizer state in client updates (Karimireddy et al., 2020) [2]. Techniques for adaptive optimization approaches for local updates in federated learning are introduced by LocalAdaptivity (FedLocal) (Wang et al., 2021b) [6]. Using theoretical input-agnostic assurances, our method, FedDA, validates the group fairness of classifiers, removing the requirement to understand the change in sensitive attributes between training and deployment datasets.

Interestingly, out of the nine baselines we tested, some use momentum just in one direction (FedLocal), while others use momentum both ways (without momentum aggregation or with aggregation but restarting momentum at each FL round). By comparison, our approach results in bigger oscillation but faster convergence since it preserves momentum aggregation throughout the FL process.

3 Methodology

Dataset

We employ three well-known tasks from the fields of natural language processing and computer vision on typical benchmark datasets. We use a CNN intended for character recognition to carry out image classification using MNIST [7]. We split the non IID approach in a ratio of 60-40, with 42000 images for training and 28000 for testing. For IID data have 600 training images and 350 testing images, which is also split 60-40. It contains images of 28x28 pixels of handwritten digits, with non IID data having evenly distributed probabilities of the same digits and IID being unevenly or randomly distributed.

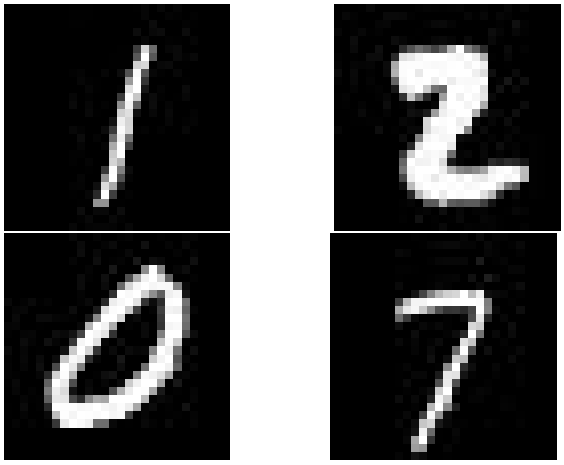

Fig. 1: Samples Images from Dataset

Baselines

A data-sharing approach is suggested which creates low sample data that is universally shared among every peripheral node in order to increase FedAvg with non-IID data. By marginally lowering EMD, we can greatly improve test accuracy for severely skewed non-IID data. We can provide the clients with a limited portion of the global data that is uniformly distributed among classes from the cloud, since we do not have control over the data of the clients. This makes sense in a normal federated learning setting during the initialization phase. Furthermore, we can prepare the model by training on the universally shared data and sent to client nodes in place of a model with random weights. Test accuracy is anticipated to increase as a result of the internationally shared data's ability to lower EMD for clients.

We used the data-sharing strategy that is demonstrated in the context of federated learning. A globally shared dataset S that is uniformly distributed across classes is centrally located in the cloud. A random α chunk of S and the preparation model that was pretrained with S during FedAvg initialization phase are sent to each client.

Every client's local model is trained using both their individual private data and the shared data from S. After that, the cloud uses FedAvg to train a global model by combining the local models from each client. We can compromise between two things: (a) between α and accuracy, and (b) between accuracy and S size, which can be expressed as $\beta = ||\mathrm{S}||\ ||\mathrm{D}|| \times 100\%$, where D is the total amount of data from the edge computers.

D is divided into ten customers using non-IID, one-class data. Ten random Ss with β ranging from 2.5% to 25% are generated using H. Initially, FedAvg merges all of the S's data with the client's data, and then trains ten CNNs on the combined data over a 300-round communication cycle. In Figure 7, the test

accuracy is shown versus β. Second, we designate two distinct S's: $G_{10\%}$ for β = 10% and $G_{20\%}$ for β = 20%; this is done. For every S, the following steps are taken: (a) train a warm-up CNN model on S to achieve a test accuracy of $\tilde{6}0\%$; (b) merge the data of each client with a random α fraction, then train the warm-up model on the merged data. In Figure 7, the test accuracy is shown versus α. The training parameters utilized for Section 3 remain unchanged. As β increases, the test accuracy rises to 78.72%, as Figure 7(a) illustrates. We can attain a test accuracy of 74.12% for the extreme 1-class non-IID data, even with a lower β = 10%, as opposed to 44% without the data-sharing technique.
It also turns out that in order to obtain a comparable accuracy, the consumers do not need to receive the complete S. Rather, each client just needs to receive a random portion of S. The test accuracy of the warm-up model steadily rises with α, as Figure 7(b) illustrates, reaching 77.08% for $S_{20\%}$ and 73.12% for $S_{10\%}$. Specifically, for both $S_{20\%}$ and $S_{10\%}$, the test accuracy increases by less than 1% when α rises from 50% to 100%, following an initially rapid climb. With a suitable selection of α, we may thereby further minimize the size of the data that is actually received by each client.
As federated learning is established, the strategy only needs to be executed once, therefore communication overhead is not a significant issue. The globally shared data is not sensitive to privacy because it is a different dataset from the clients' data.

4 Experiments

Using the MNIST dataset, we ran experiments in this paper to assess the effectiveness of federated learning models under two different data distribution scenarios: non-identically distributed (Non-IID) and independently distributed (IID). The MNIST dataset, which consists of handwritten digit pictures, was used as an example case study to investigate how heterogeneity in data distribution affects federated learning.

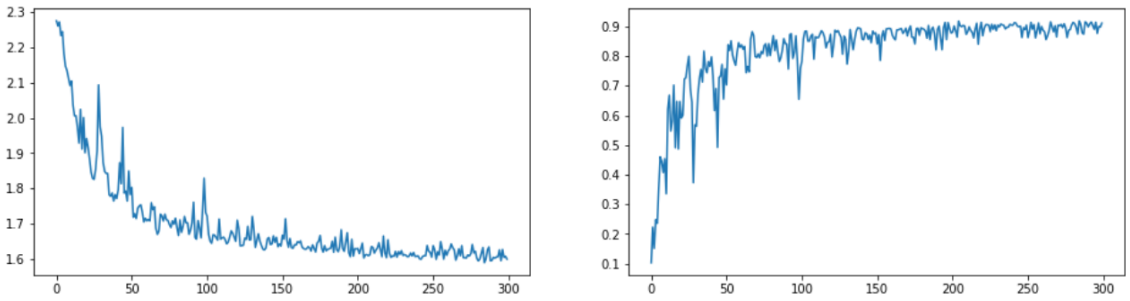

Fig. 2: Loss and Accuracy Graph On IID dataset

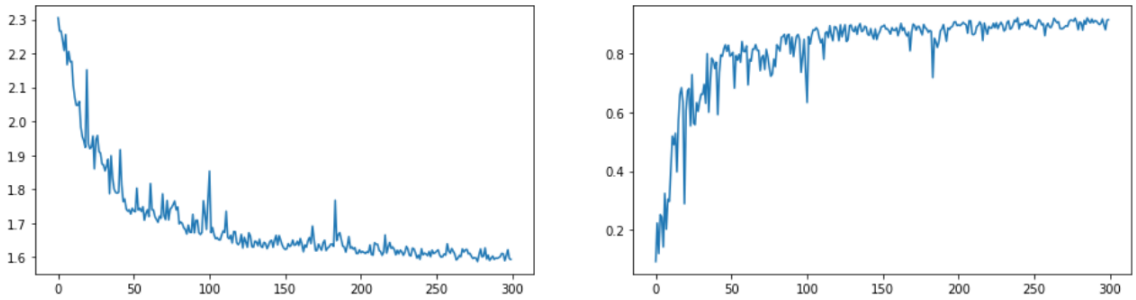

Fig. 3: Loss and Accuracy Graph on Non-IID data

For the IID setting, we randomly partitioned the MNIST dataset into subsets, ensuring that each subset had a similar distribution of handwritten digit images. Conversely, for the Non-IID setting, we intentionally created non-uniform distributions across the subsets, simulating scenarios where edge devices possess different distributions of handwritten digit images.
We employed Federated Learning methodologies to train neural network models on decentralized devices, simulating a real-world federated learning environment. The training process involved iterative model updates and parameter exchanges between the central server and the participating devices, while preserving data privacy on the edge devices. To assess the performance of the Federated Learning models under varying data distributions, we measured the accuracy and loss on both IID and Non-IID data settings.

5 Results Analysis

The findings from our experimentation underscore the impressive adaptability of Federated Learning models in the face of diverse data distributions across decentralized devices. After rigorous testing, our model achieves a conclusive accuracy of 87.69% and a loss of approximately 1.5996% on Independently and Identically Distributed (IID) data. In the realm of non-IID data, our model exhibits an even more noteworthy accuracy of 91.6%, coupled with a loss of roughly 1.5930%.
Furthermore, our research delves into the intricate repercussions of data distribution heterogeneity on Federated Learning. We reveal that neural networks trained on datasets exhibiting pronounced skewness and non-identical distribution patterns experience a notable decline in Federated Learning accuracy, with an estimated reduction of around 5%.
The broader implications arising from this investigation shed light on the pivotal importance of comprehending data distribution dynamics in Federated Learning environments. These revelations offer indispensable insights for refining Federated Learning methodologies, particularly in navigating the challenges presented by data heterogeneity. By addressing these challenges proactively, our study not

only contributes to augmenting the adaptability of models but also identifies promising avenues for enhancing the scalability of decentralized learning systems in real-world applications.
In essence, a profound understanding of the intricacies of data distributions in Federated Learning emerges as a critical factor in advancing the overall efficacy and applicability of decentralized learning paradigms.

6 Conclusion

The study's findings underscore the significance of Non-Independent and Non-Identically Distributed (Non-IID) data in Federated Learning, particularly in real-world applications where data distributions across decentralized devices are diverse and independent. The observed higher accuracy of the Federated Learning model in Non-IID data settings, as compared to Independent and Identically Distributed (IID) scenarios, aligns with the characteristics of real-world data environments. This resemblance to real-world data distribution highlights the model's adaptability to varying data distributions across decentralized devices, emphasizing its robustness in handling data heterogeneity.
Furthermore, the implications of these findings extend to the enhancement of model adaptability in real-world applications, where data privacy and diverse data distributions are paramount. Understanding the nuances of Non-IID data distributions in Federated Learning environments is crucial for advancing the efficacy and scalability of decentralized learning systems.

7 Future Work

The study only evaluates the performance of Federated Learning models using the MNIST dataset, which is a relatively simple dataset. The results may not generalize to more complex datasets, and further research is needed to evaluate the performance of Federated Learning models on more challenging datasets. Extending model training to larger and more diverse datasets, such as CIFAR-100 with its 100 classes, presents an opportunity to evaluate federated learning approaches on complex data landscapes. Additionally, exploring different datasets beyond CIFAR-100 will help assess FL models' performance and generalization across various domains.
Exploring innovative methodologies to enhance various aspects of federated learning, such as communication efficiency, convergence speed, optimization robustness, and privacy preservation, is essential. These efforts aim to make FL algorithms more efficient, reliable, and scalable for broader deployment. We will also use other models such as FedTL , A federated learning model explicitly designed for transfer learning, FedProx, a model that introduces a proximal term to the optimization objective which promotes more robust convergence in non-IID scenarios, and FedDyn, which explores dynamic strategies for federated learning, dynamically adjusting model parameters to adapt to changing non-IID data distributions.

References

1. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
2. S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.
3. A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani, "Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 606–14 619, 2021.
4. Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, 2020.
5. S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.
6. J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data *et al.*, "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.
7. L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.