Federated Learning on Non-IID Data

Project Plan

Wong Ngai Sum (3035380875) Project Supervisor: Dr. Heming Cui

1 Introduction

As technology advances, processor chips become increasingly powerful and mobile processors are as fast as most desktop computers. Smartphones and IoT devices are everywhere and a large amount of data are being collected through sensors for analytics. They generate enormous valuable data and machine learning is applied to those data to improve the intelligence of services. Standard machine learning requires operations in data centers with a centralized dataset to attain fast processing and low latency for transmission. However, it comes with several constraints such as data privacy and the inability to process large dataset. Serverless architecture and distributed architecture have become ways for many developers to have more flexibility and improve their services. *Federated Learning*, a new decentralized AI framework presented by Google in 2015 [1], enables IoT and mobile devices to collaboratively learn a shared machine learning model, while reducing privacy and security risks, attaining lower latency and providing personalized models to clients. Federated Learning has been tested in Gboard on Android, which stores information about the current context to train its query suggestion model [2]. There are some Federated Learning frameworks available now, such as Tensorflow Federated, FATE, and PySyft.

Federated Learning has several key properties that differentiate from typical distributed edge learning, including Non-IID (highly skewed) data, varying amounts of training data between clients, massive distribution of data and limited internet connection. Specifically, it has issues with serious communication overhead [3][4][5], low accuracy on Non-IID data [5] and privacy attacks [6]. Some research has proposed several techniques to address these problems, such as subsampling, probabilistic quantization [5] and globally shared data [6]. However, there is large room for improvement to make Federated Learning more efficient and mainstream [8].

This project mainly focuses on issues with Non-IID data. Experiments will be carried out and explanations of these phenomena will be given. I also intend to give some suggestions to heighten accuracy with Non-IID data, while not adding much communication costs.

The following of this proposal is structured as follows: Section 2 covers the background of Federated Learning and non-IID data; Section 3 describes the scope and deliverable of the project; Section 4 presents the intended methodology to be applied in the project; Section 5 discusses the expected challenges; Section 6 presents project schedule and Section 7 concludes.

2 Background

2.1 Non-IID Data

First, we need to know what is IID data. IID data means independent and identically distributed data. Independent means the sample items are not connected in any way. Identically distributed means all sample items are taken from the same probability distribution. Non-IID data is very common, such as search queries entered by a guy and his forum posts, because they may be related to his background and the wordings are similar. If he is a mathematician, most of his data may be related to mathematics. Non-IDD data noise is a common challenge in many disciplines including neuroscience [12]. It may cause biased estimation in Federated Learning due to weight divergence.

2.2 Federated Learning

Federated Learning can be divided into 4 steps [3][4]: i) clients download parameters of a training model from a server; ii) a client updates the model with their local data; iii) upload the new model parameters to the server while keeping all the training data on device; iv) server aggregates the updates to improve the model. Due to the slow and unstable network of mobile devices, *Federated Averaging* algorithm is typically used to train neural networks using 10-100x less communication than federated SGD (Stochastic Gradient Descent) [2]. However, FedAvg algorithm reduces accuracy significantly with highly skewed non-IID data, up to 11% for MNIST, 51% for CIFAR-10 and 55% for keyword spotting datasets [6]. Improving accuracy with non-IID is still a main challenge for many researchers.

2.3 Literature review

One paper has been discussed here related to this project, “Federated Learning with Non-IID Data” [6]. This paper shows that the accuracy reduction is less for the two-class non-IID data than for one-class non-IID data, so the accuracy of Federated Averaging algorithm will be affected by the data distribution. The accuracy reduction can also be explained by the weight divergence, which can be evaluated by the earth mover’s distance between the distributions. To tackle the problem of weight divergence, globally shared data are used to train a warm-up CNN model to test accuracy of ~60%, then it is distributed to each client with a small subset of data. As a result, accuracy can be increased by ~30% for the CIFAR-10 dataset with only 5% global shared data.

However, this paper only tested the data-sharing technique in one use case and assume the amount of data distributed to each client is equal. Also, it does not explain why a warm-up model should be trained on shared data and why test accuracy should be trained to ~60%. More investigations can be made to evaluate the effectiveness of this technique.

3 Objectives

3.1 Scope

Due to the time constraint of this project, the scope of this project is tentative and subject to changes.

Trade-off analysis. There are lots of learning parameters that may affect accuracy, communication costs, and convergence speed. It is crucial to figure out relationships between parameters. Moreover, several parameters need to be fine-tuned in different use cases [6], so optimal values for most use cases will be tested.

Accuracy analysis. There are other machine learning architectures, such as centralized training, large-batch training, and distributed training in data centers. There are also different data distributions, such as uniform distribution, one class per device, two classes per device and so on. Some clients may have more or fewer data than others in real life. The recent paper proposed globally shared IID data and warm-up model to improve accuracy by ~30% with only 5% shared data [6]. However, it may not reflect real-world scenarios as amounts of data on each device are assumed identical and different types of learning tasks have not been tested. Several experiments with different parameters will, therefore, be done to analyze the results of accuracy and biased estimation.

Theory of poor accuracy with non-IID data. Federated Learning is suitable for IID data instead of non-IID data [4]. The primary goal of this project is to explain why Federated Learning with non-IID data has poor accuracy. It can be done by analyzing test results and the difference from other architectures. A theory will be proposed to explain the phenomena.

Communication-efficient federated optimization techniques. The main idea is to improve accuracy while not affecting speed and communication frequency too much. There are two intuitive ways to increase computation: i) increased parallelism and ii) increased computation on clients [4]. It is interesting to test how increasing and decreasing computation will affect accuracy and convergence speed with non-IID data. More papers will be read to incorporate their optimization methods. Experiments will also be done to prove and explain why these suggestions are effective.

3.2 Deliverable

This project will deliver: i) a trade-off analysis for different learning parameters like weight divergence and accuracy [6]; ii) an analysis of test accuracy with Federated Learning on non-IID data, as well as results with other learning architectures (e.g. Large-Batch Training and Centralized Training) or IID data; iii) a theory to explain why Federated Learning on non-IID data has poor accuracy; iv) potential communication-efficient federated optimization techniques.

4 Methodology

To improve the reliability and validity of this project, three main phases are divided and the methodology of each phase are as follows:

Evaluations. I will conduct some controlled experiments similar to real-life use cases, such as word prediction, image classification, and data partitioning. I may try to present datasets like MNIST [9], VGG[10] and CIFAR-10 [11]. If some experiments on research papers seem good, I may reimplement them and try with different parameters. I will try different values for parameters to do trade-off analysis. Evaluations on different architectures will be done and different data distributions will be tested. Amounts of data on each client may be different or the same.

Explanations. Experimental results will be presented and analyzed. I will elaborate on the description of tests as well as the processes to gather and analyze data using quantitative methods. As a large amount of information will be presented, the information will be organized sub-sections.

Suggestions. I will read more papers to see what techniques were proposed to improve accuracy, training speed and communication overhead with non-IID data. After doing the literature review, I can know what obstacles they are facing to make Federated Learning mainstream. I may come up with several approaches to improve Federated Learning on non-IID data. In case these suggestions fail, analysis can be done to explain the reasons behind.

5 Challenges and Mitigation

I am expecting two main challenges in this project:

- Very limited time and personal
  I will prepare ahead of time. In case it is impossible to complete all desired tasks within the timeframe, time will be spent on fulfilling the primary goals of this project.

- Unfamiliar with Federated Learning frameworks and cluster deployments
  I will read their documentation and search for other online resources. I will also request help from the supervisor's Ph.D. students to set up test environments.

6 Project Schedule and Milestones

Periods	Milestones
September	● Literature review ● Analysis of existing algorithms ● Study Federated Learning frameworks
October	● Literature review ● Write testing functions ● Experimentation
November	● Suggestions ● Results Analysis and Comparison

December	• Final report writing • Project presentation

7 Conclusion

This is a detailed plan for this project. Federated Learning is a decentralized collaborative machine learning approach that provides better privacy, less latency, and smarter model. Unfortunately, it also comes with communication efficiency and data accuracy problem. While some effort has been made on improving accuracy with non-IID data, further research can be done by incorporating their experiments and optimization methods and performing trade-off analysis. Therefore, this project intends to propose a theory to explain the experimental results, and hopefully, suggest some communication-efficient federated optimization techniques.

8 References

[1] Konečný, J., McMahan, B., & Ramage, D. (2015). Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*.

[2] Federated Learning: Collaborative Machine Learning without Centralized Training Data. (2017, April 6). Retrieved from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[3] Nishio, T., & Yonetani, R. (2019, May). Client selection for federated learning with heterogeneous resources in mobile edge. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1-7). IEEE.

[4] McMahan, H. B., Moore, E., Ramage, D., & Hampson, S. (2016). Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*.

[5] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[6] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

[7] Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

[8] Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data. *arXiv preprint arXiv:1903.02891*.

[9] "THE MNIST DATABASE." *MNIST Handwritten Digit Database, Yann LeCun, Corinna Cortes and Chris Burges*, http://yann.lecun.com/exdb/mnist/.

[10] *VGG Face Dataset*, http://www.robots.ox.ac.uk/~vgg/data/vgg_face/.

[11] *CIFAR-10 and CIFAR-100 Datasets*, https://www.cs.toronto.edu/~kriz/cifar.html.

[12] Georgiev, K., & Nakov, P. (2013, February). A non-iid framework for collaborative filtering with restricted boltzmann machines. In *International conference on machine learning* (pp. 1148-1156).