Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230721

Battery-aware federated learning: Challenges and solutions

Ziyue Zhang
Wuhan Britain-China School, Wuhan, China

twu43@student.ubc.ca

Abstract. Smartphone battery life is a pivotal factor in consumers' purchasing decisions. Recent years have witnessed a surge in studies focusing on smartphone energy management, with data-driven energy management systems offering solutions to prolong battery life. Federated Learning (FL) emerges as a promising distributed learning algorithm, enabling wireless devices to upload locally trained models, fostering collaborative learning without exposing sensitive data. This paper explores the FL process, particularly the Federated Averaging (FedAvg) approach, which excels in scenarios with homogeneous data. In the era of burgeoning data generation, traditional cloud computing systems face limitations, driving the adoption of Edge Computing (EC), which processes data closer to its source, enhancing response times. To make FL efficient for e-commerce, resource constraints must be addressed. This involves techniques like local updates and model compression, which reduce communication overhead. However, FL brings challenges related to data distribution heterogeneity and privacy concerns. Solutions like differential privacy, encryption, and access control are discussed. In conclusion, this paper presents an overview of smartphone battery life, data-driven energy management, and the potential of FL, emphasizing its relevance in the age of EC. By addressing resource limitations and privacy issues, FL holds promise for efficient data processing.

Keywords: The Federated Learning, Battery Life Extension, FedAvg.

1. Introduction

Smartphone battery life is a crucial factor that consumers consider when making a purchase. In recent years, numerous studies have focused on managing smartphone energy usage [1]. Thanks to the development of data-driven energy management systems, there are now many solutions available to extend the lifespan of smartphone batteries.

The emerging federated learning (FL) architecture represents one of the most promising distributed learning algorithms [2]. In the domain of federated learning, wireless devices no longer need to share their entire training data with a central entity. Instead, they can upload their locally trained learning models to a base station, facilitating collaborative learning [3]. With model training taking place at the edge of the network, direct data exchange is no longer required. FL's data decentralization promotes effective and secure collaborative learning, allowing wireless devices to contribute unique insights without compromising data privacy.

The FL process enables clients to benefit from collective knowledge without compromising data privacy. In this study, the federated averaging (FedAvg) approach was investigated, which is the standard for FL in many other studies. FedAvg is well-suited for situations with homogeneous data across clients. Numerous local updates can enhance model performance on all clients' data when the

DOI: 10.54254/2755-2721/41/20230721

data is similar. Updates from one client can improve the model's performance on other clients' data, as long as the data exhibits similarities.

In today's digital world, the generation and processing of data are growing at an unprecedented rate. Consequently, traditional cloud computing systems are becoming increasingly inadequate for handling the massive volume of data. This is where Edge Computing (EC) comes into play [4]. EC, a distributed computing paradigm, involves storing data locally and migrating high-computing power applications to the network's edge. Data processing occurs closer to the data source, reducing latency and improving response times.

However, in developing efficient FL for e-commerce, it's essential to address resource constraints. The limitations of FL resources have been discussed in earlier publications. In machine learning, training models often require substantial amounts of data, making data transmission time-consuming.

To tackle this challenge, researchers have devised techniques to perform model training through numerous local updates before global aggregation. The concept behind this approach involves dividing data into smaller subsets and conducting local training on each subset. This minimizes the data that needs to be transmitted across the network, as only model updates must be shared among devices. Once all local updates are collected, they can be globally aggregated to produce a final model that represents the entire dataset. Additionally, model compression technology [5] is frequently used to further reduce communication overhead. Techniques such as quantization [6] or sparsification [7] are employed to achieve effective model aggregation.

While FL offers numerous benefits, it also presents challenges. The heterogeneity of data distribution among multiple partners poses a significant obstacle to federated learning [8]. Data may not be evenly distributed among parties in reality, which can affect the effectiveness of federated learning [9]. When parties modify their local models, their objectives may differ substantially from global goals. Consequently, the global average model may fall short of being globally optimal.

Furthermore, FL raises privacy concerns, as data processing and storage occur locally. These devices may store personal data, which could be exposed or leaked during local data model training. Additionally, model updates are shared between the centralized computer and participant devices at various points during the federated learning process. This data transmission increases the risk of data interception or tracking, compromising user privacy. Moreover, if aggregation methods do not prioritize privacy, they may inadvertently reveal information about specific data points or individuals. Attackers may attempt to extract valuable information or scrape local data from the aggregated model.

In response to these challenges, several potential solutions are discussed in Section 4.

In summary, this paper has explored various aspects of smartphone battery life, data-driven energy management systems, and the promising field of federated learning. It has highlighted both the advantages and challenges of federated learning and its relevance to Edge Computing. By addressing resource constraints and privacy concerns, we can harness the potential of federated learning for efficient data processing in today's data-driven world.

2. Federated Learning

2.1. Implementation principle

Typically, federated learning is implemented by offloading computing tasks to more powerful servers and cloud platforms to extend the life of smartphones. Rather than simply relying on the limited capabilities of smartphones, federated learning allows devices to connect to remote servers or cloud infrastructure so complex calculations can be performed. This reduces the load on the smartphone hardware, extending battery life and improving reliability. FedAvg is the name of the traditional federated learning method. Its basic concept is shown in Figure 1. The following steps are used to implement the FedAvg method:

1) Each device downloads a common global model for subsequent local training.

DOI: 10.54254/2755-2721/41/20230721

2) Use local data from different mobile devices to improve the entire download model through multiple independent local updates. The corresponding gradient information is then uploaded to the cloud in encrypted form.

3) As an average update of the local model published to the cloud, the new global form is provided to the device.

4) Repeat the above steps until the model works as expected or until the deadline expires.

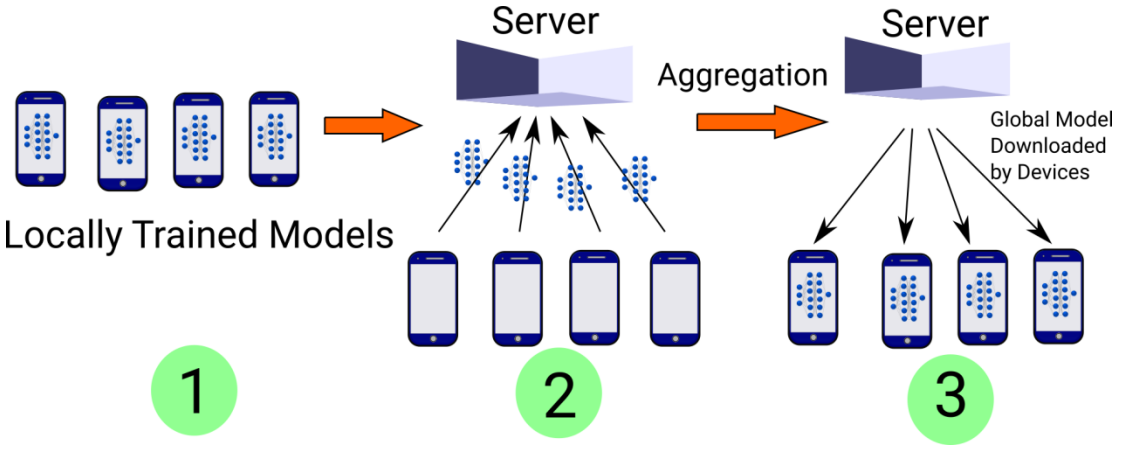

Figure 1. Basic Concept of FedAvg [10].

3. How does Federated Learning extend battery lifetime

To extend the life of your smartphone battery, reducing battery usage is crucial. The communication function of a smartphone is one of the most power-consuming functions. Therefore, the learning method may be used with two fundamental message compression techniques to decrease the consumption of communication energy: local update and model compression. The former does not transfer raw data to the server, but updates the local model with its own data. The latter significantly reduces the amount of bits required to transmit model changes by using compression techniques such as quantization or sparsification, thereby reducing the communication payload in each cycle.

3.1. Local updating

3.1.1. Working principle

Local updating ensures that private data is kept private by allowing each participating device or organization to independently train a machine learning model using its own data. Local training of the model allows the device or organization to better understand and incorporate the specific patterns and insights found in its data, improving model performance and accuracy.

Collaborative learning is made possible by the consolidation of several local model updates because the collective intelligence of all participants is used to enhance the performance of the model as a whole. Importantly, this strategy does away with the need to exchange or keep sensitive data centrally, protecting privacy and guaranteeing data security. Federated learning enables companies to cooperate and gain from the network's collective intelligence while preserving strict privacy requirements by keeping data decentralized.

3.1.2. Stochastic Gradient Descent (SGD)

Model training is made easier by Stochastic Gradient Descent (SGD), a popular optimization approach in machine learning that iteratively modifies the model's parameters to reduce the loss function. SGD presents a random sampling method termed mini-batch selection in contrast to standard Gradient Descent, which computes the gradient using the whole dataset. In order to calculate the gradient and update the parameters using this approach, a subset (mini-batch) of the data must be chosen, adding

DOI: 10.54254/2755-2721/41/20230721

randomization to the procedure. SGD is especially useful when working with huge datasets since it decreases the computing strain by taking into account only a tiny percentage of the data in each iteration. This randomized method enables quicker and more efficient updates. The stochastic aspect of SGD also helps the algorithm to traverse beyond local minima, examining a wider variety of solutions and improving the model's ability to generalize to unobserved data. By providing computing efficiency, increased generalization, and a way to avoid less-than-ideal outcomes, SGD therefore plays a crucial part in improving machine learning models.

3.1.3. Reducing the total number of communication rounds

Communication rounds refers to the number of times a client device communicates with a server in federated learning. The more communication rounds, the higher the communication cost and the lower the system efficiency. Therefore, reducing the number of communication rounds is essential to improve the scalability and efficiency of federated learning.

Prior to the deployment of the gradient, the number of local modifications made by each client device directly affects the number of communication rounds. This indicates that additional communication rounds are needed to finish the model training if each client device only makes a modest number of local modifications. Therefore, the overall number of communication cycles can be decreased by carrying out local updates in phases. On each client device, batch local updates are gathered and then communicated collectively to the server. By doing so, the system's effectiveness and the number of communication cycles may both be increased.

3.2. Model Compression

Model compression, which entails quantization and sparsification, has also been investigated as a means of minimizing the volume of messages transferred between devices or servers during distributed learning, in addition to local updating options. While sparsification entails splitting up big messages into smaller ones that may be conveyed more effectively, quantization involves decreasing the accuracy of the data being communicated. In early dispersed learning research, both conceptual and empirical studies of these techniques have been conducted. These model compression techniques face new challenges due to local update systems, non-identically distributed local data, and low device participation in federated settings. In federated environments, many publications present useful strategies such as requiring low-level, non-common update models [11], using structured stochastic rotation for quantitative analysis [11], and reducing the number of servers and device-to-device communications [12].

3.2.1. Gradient Sparsification

In FL, communication and computation must be balanced to reduce the energy required to train the model. This trade-off is modified in the original FL strategy, Federated Avg (FedAvg) [13]. FedAvg passes either all model parameters or no parameters after each local gradient descent model update step.

Gradient sparsification (GS) is a technique used to reduce the amount of data transferred during the update of a machine learning model. This approach involves sending only sparse vectors that contain a subset of large values from the full gradient. By doing so, a more balanced approach is achieved, which can significantly improve communication efficiency. The process of GS involves compressing the gradient or weight vector of the model, which is done by selecting only the most significant values and ignoring the rest. This compression reduces the amount of data that needs to be transferred, which is critical when working with large datasets and distributed computing systems. The benefits of using GS are numerous. By reducing the amount of data transferred, GS can help reduce the cost of computing and improve the speed of model training. Additionally, it can help overcome bandwidth limitations and enable more efficient communication between the nodes in a distributed system. The technique of GS is particularly useful in scenarios where the gradient or weight vector is too large to be transferred efficiently. In such cases, compressing the vector can significantly improve the efficiency of the system. Moreover, GS can be used in conjunction with other techniques such as quantization, which further reduces the amount of data transferred during communication.

DOI: 10.54254/2755-2721/41/20230721

3.2.2. Gradient Quantization

Deep learning technology, which enables machines to learn from data, has led to the creation of larger and more complex models. However, as the models become more complex, they require increasing amounts of computational resources to run effectively. The challenge of deep learning is the increasing computational resources required. As Deep learning models are often trained on large datasets, which requires a significant amount of processing power. Another challenge of deep learning is the data transmission overhead. As models become larger, they require more data to be transmitted between nodes during training. This can lead to significant delays and increased costs, especially when training on a distributed system. To solve this problem, researchers have proposed many compression methods, one of which is gradient quantization. Gradient quantization is similar to gradient dilution and is an effective compression method that can reduce energy consumption during transmission without affecting model performance.

After the local training of a single epoch is completed, the calculated gradients need to be submitted to the aggregator for global update. However, due to limitations in computing power, the accuracy of local gradients may not be high, or fewer bits may be used. To reduce the cost of communication per round, gradient quantization allows us to quantize these local gradients instead of loading all raw gradients. In this way, high-level gradient quantization at each round can be guaranteed with minimal communication energy usage.

However, the reduced accuracy requires more global rounds for correction, which results in more computer power being consumed overall. Therefore, in order to find a balance between communication and processing and improve overall energy efficiency, quantization level should be adjusted. This can be achieved by dynamically changing the gradient quantization level to minimize the use of communication and computing resources while maintaining model performance.

4. Challenges

4.1. Statistical Heterogeneity

In a meta-analysis or systematic review, statistical heterogeneity refers to the variation or variety in effect sizes or outcomes seen within individual studies. This suggests that there are differences in the findings or conclusions of the included research that go beyond what would be predicted by chance.

Different research designs, demographic characteristics, interventions or treatments, outcome measures, or other sources of variance can all lead to statistical heterogeneity. It is often measured using statistical techniques, with the I2 statistic calculation being the most used technique. The I2 statistic shows what percentage of the effect size's overall fluctuation is attributable to heterogeneity rather than random error.

High statistical heterogeneity suggests that the investigated studies are not sufficiently consistent or comparable to warrant combining their findings in a meta-analysis. In this situation, investigating potential sources of heterogeneity or doing subgroup analyses may be more suitable in order to comprehend the causes of variability and offer context for the study's findings.

4.2. Privacy concerns

Data exposure is the initial privacy concern with federated learning. Data leaks during model updates are a possibility since training occurs on a local device or server. Attackers could try to harvest private information from these updates, violating users' privacy. Data protection techniques like encryption can help to lower this danger. Sensitive information is safeguarded during the training process by encrypting the data so that it cannot be viewed by unauthorized parties.

The ability for information to be extracted at the person level presents another privacy concern with federated learning. An attacker may still be able to obtain personal information through analytical model upgrades even if the data is encrypted. This may result in privacy breaches and jeopardize the confidentiality of data suppliers.

DOI: 10.54254/2755-2721/41/20230721

Furthermore, privacy issues may arise from participants in federated learning disclosing personal information to one another. In some circumstances, participants may be required to exchange data in order to collaboratively determine model changes. However, there is a chance that this data sharing will be accessed or used improperly.

In conclusion, federated learning raises privacy issues such as data exposure, extraction of personal information, sharing of private data, illegal access, re-identification hazards, and potential privacy breaches. private.

5. Possible Solutions

5.1. Statistical heterogeneity solution

A variety of approaches may be used to address statistical heterogeneity, which is the diversity of research findings within a meta-analysis or systematic review. Finding possible sources of heterogeneity, such as variations in research design, participant characteristics, or therapies, is crucial in the beginning. If heterogeneity is brought on by certain subgroups, subgroup analysis can be used to categorize research based on pertinent traits. Sensitivity analyses can then be carried out to determine how each study's influence on the final results will be measured. Conducting sensitivity analyses is crucial throughout this procedure to evaluate the validity of the study's findings. The I2 statistic and other reporting metrics of heterogeneity can be used to estimate the degree of heterogeneity. Instead of relying exclusively on a single pooled estimate where heterogeneity remains, it may be preferable to publish subgroup-specific effect estimates. In the end, these methods seek to increase our comprehension of the causes of heterogeneity and offer a more thorough and trustworthy justification for study findings.

5.2. Privacy solutions

There are several strategies that may be applied to overcome privacy issues in federated learning. While differential privacy offers an additional degree of protection by providing noise to prohibit information extraction at the individual level, encryption ensures that sensitive data is safeguarded throughout training. Participants can calculate model changes jointly via secure multi-party computing without disclosing their personal information. In federated learning, strict access control and data anonymization techniques reduce unwanted access and the possibility of re-identification. Continuous audits and monitoring can assist find privacy violations so that timely remedial action can be performed. Federated learning strikes a balance between safeguarding data privacy and developing collaborative models with the help of these privacy-enhancing methods.

6. Conclusion

This article serves as an example of how federated learning may be used. FedAvg is a collaborative approach of FL that enables several clients or servers to work together towards training a global model. The primary objective of this approach is to leverage the collective intelligence of multiple devices while ensuring data privacy and security. Additionally, strategies for extending smartphone batteries in terms of local updating and model compression were covered. The concerns of statistical heterogeneity and secrecy, which also have a number of potential solutions, are the primary difficulties associated with the implementation of FL.

References

[1] E. Kim, H. Jeong, J. Yang and M. Song, “Balancing energy use against video quality in mobile devices,” in IEEE Transactions on Consumer Electronics, vol. 60, no. 3, pp. 517-524, Aug. 2014.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, “Communication-efficient learning of deep networks from decentralized data,” 2016, arXiv:1602.05629. [Online]. Available: http://arxiv.org/abs/1602.05629

DOI: 10.54254/2755-2721/41/20230721

[3] J. Koneˇcný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on device intelligence," 2016, arXiv:1610.02527. [Online]. Available: http://arxiv.org/abs/1610.02527

[4] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.

[5] W. Y. B. Lim et al., "Federated learning in mobile edge networks: A comprehensive survey," IEEE Commun. Surv. Tuts., vol. 22, no. 3, pp. 2031–2063, Jul.–Sep. 2020.

[6] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized SGD and its applications to large-scale distributed optimization," in Proc. Int. Conf. Mach. Learn., 2018, pp. 5325–5333.

[7] X. Sun, X. Ren, S. Ma, and H. Wang, "meProp: Sparsified back propagation for accelerated deep learning with reduced over-fitting," in Proc. Int. Conf. Mach. Learn., 2017, pp. 3299–3308.

[8] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977, 2019.

[9] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for ondevice federated learning. In Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020.

[10] Aniruddh Herle, and Janamejaya Channegowda. "Federated Learning: An Energy Efficiency Perspective to Extend Smartphone Battery Life." 2022 IEEE Delhi Section Conference (DELCON), 11 Feb. 2022, https://doi.org/10.1109/delcon54057.2022.9753113. Accessed 2 Sept. 2023.

[11] J. Konecˇny`, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, Federated learning: Strategies for improving communication efficiency. 2016. [Online]. Available: arXiv:1610.05492.

[12] S. Caldas, J. Konecˇny, H. B. McMahan, and A. Talwalkar, Expanding the reach of federated learning by reducing client resource requirements. 2018. [Online]. Available: arXiv:1812.07210

[13] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in AISTATS, 2017.