FEDERATED LEARNING FOR PRIVACY PRESERVATION

SUBMITTED BY

Group Number: 10

MITODRU GHOSH, 200010938917, 2011200001093

IPSITA SANNYASHI, 200010742209, 2012200001074

TANMOY DAS, 200010637453, 2011200001063

SOURADEEP DE, 200010597288, 2011200001059

Under the Supervision of

Prof. Dr. Souvik Pal

Associate Professor

Department of Computer Science & Engineering

SISTER NIVEDITA UNIVERSITY

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to our project guide, **Prof. Dr. Souvik Pal**, Associate Professor, Department of Computer Science and Engineering, Sister Nivedita University, for his invaluable guidance, constant encouragement, and insightful critiques of this research work. His perceptive criticism kept us working to make this project in a much better way. Working under him was an extremely enlightening experience for us and we are thankful for his encouragement during the course of this project.

We are also immensely grateful to all the faculty members of the Department of Computer Science and Engineering for their direct and indirect support during the course of this project.

Last but not least, we would like to thank our team members, whose diligence, understanding, and cooperation have been a vital contribution to this project. It was a pleasure to work with them and we are proud of what we have accomplished together.

We are grateful for the learning and growth we experienced throughout this project and look forward to applying these lessons in our future endeavors. Thank you all for your unwavering support and guidance.

Content

1. INTRODUCTION

Federated Learning is a revolutionary approach to machine learning that allows for the training of algorithms across multiple devices or servers holding local data samples, without the need to exchange their data samples. This approach is particularly useful in preserving the privacy of data, a critical concern in today's data-driven world.

The project titled "Federated Learning for Privacy Preservation" aims to explore and implement this innovative machine learning approach in the context of our B. Tech Computer Science & Engineering degree. The primary objective or aim of our project is to identify and analyze pre-existing Federated Learning Algorithms currently in use and then to opt for one of the suited algorithms from the sorted ones and then to study and observe it thoroughly from a research's point of view thereafter to optimize the same algorithm for further cost and time efficiency across servers and devices from decentralized data while ensuring the privacy of individual data points.

This project will delve into the intricacies of federated learning, including its architecture, algorithms, and communication protocols. It will also address the challenges associated with this learning paradigm, such as system heterogeneity, communication efficiency, and model aggregation.

By focusing on privacy preservation, the project aligns with the growing need for data security and privacy in the era of Big Data and Artificial Intelligence. It underscores the importance of ethical considerations in machine learning and aims to contribute to the development of more secure and privacy-preserving machine learning models.

In conclusion, this project stands at the intersection of machine learning, data privacy, and distributed systems, offering a comprehensive study and practical implementation of Federated Learning for Privacy Preservation. It promises to be an exciting journey of learning and discovery in the realm of privacy-preserving machine learning.

2. METHODOLOGY:

A. **Literature Review**: Begin by conducting a comprehensive literature review to understand the current state of federated learning algorithms. This will involve studying various research papers, articles, and resources to gain a deep understanding of the existing algorithms and their performance metrics.

B. **Algorithm Selection**: Based on the literature review, identify a set of promising algorithms that are suitable for federated learning and privacy preservation. The selection should consider factors such as computational efficiency, communication cost, and privacy preservation capability.

C. **Data Preparation**: Identify or create a suitable dataset for testing the algorithms. The dataset should ideally be distributed across multiple devices or servers to mimic a real-world federated learning scenario.

D. **Algorithm Implementation**: Implement the selected algorithms using a suitable programming language or framework. Ensure that the implementation is faithful to the original algorithm design and can work with the prepared dataset.

E. **Algorithm Optimization**: Once the algorithms are implemented, the next step is to optimize them. This could involve tuning hyperparameters, modifying the algorithm structure, or even proposing new methods to enhance efficiency and reduce cost.

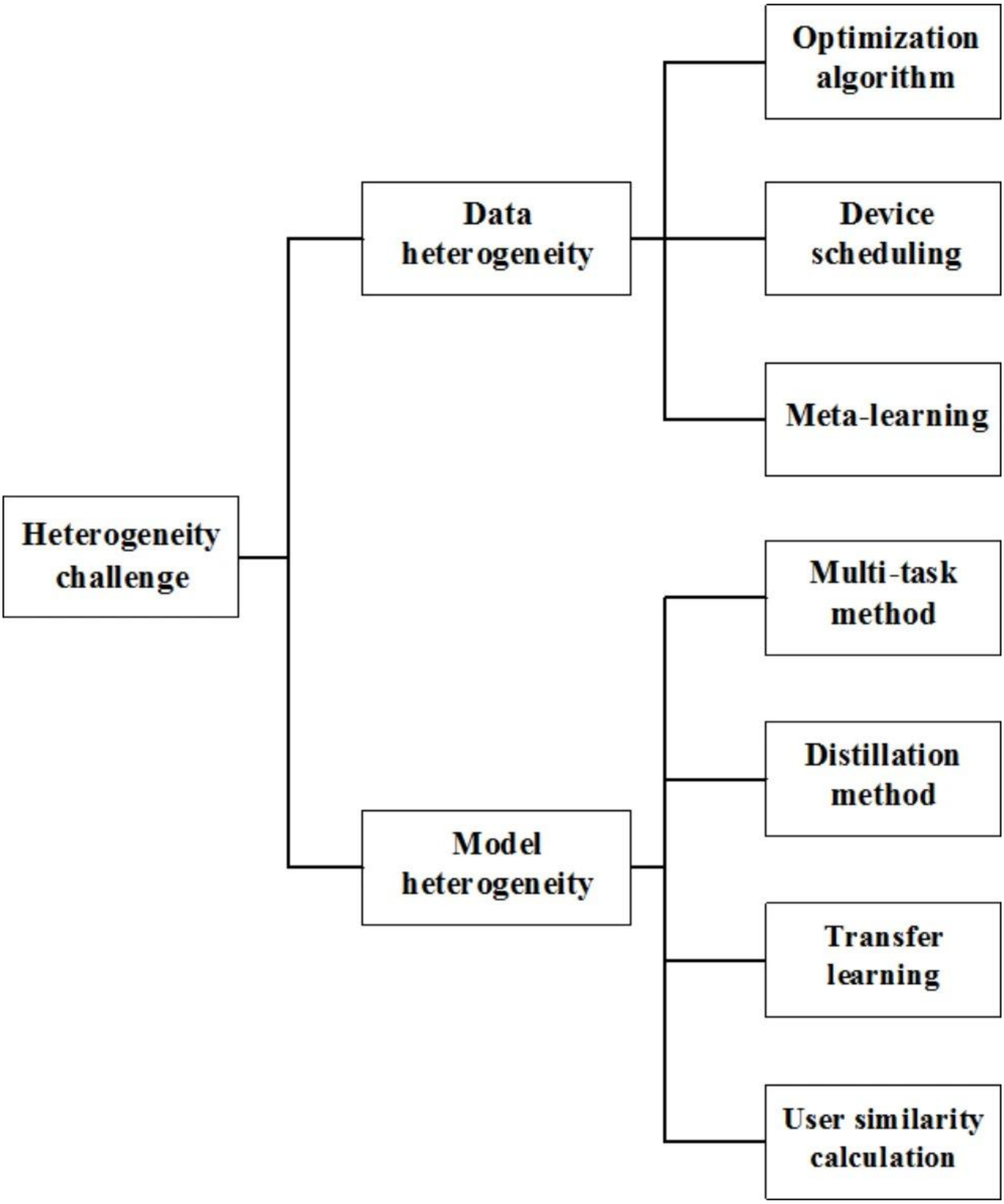
F. **Evaluation**: Evaluate the performance of the optimized algorithms. This should involve both quantitative metrics (such as model accuracy, communication cost, and computational efficiency) and qualitative analysis (such as privacy preservation capability).

G. **Comparison and Analysis**: Compare the performance of the different algorithms and analyze the results. Identify the strengths and weaknesses of each algorithm and discuss possible reasons.

H. **Documentation**: Document all the processes, findings, and conclusions in a well-structured report. The report should clearly explain the methodology, results, and implications of the findings.

I. **Presentation**: Prepare a presentation summarizing the project. The presentation should be designed to communicate the project's objectives, methodology, findings, and conclusions to a non-technical audience.

This methodology provides a systematic approach to achieving the project's objectives and contributes to the field of federated learning and privacy preservation. It's important to note that this is a general methodology and may need to be adapted based on the specific requirements and constraints of our project.


Heterogeneity challenge
Data heterogeneity
Model heterogeneity
Optimization algorithm
Device scheduling
Meta-learning
Multi-task method
Distillation method
Transfer learning
User similarity calculation

3. PROBLEM SPECIFICATION

The problem specification for the project "Federated Learning for Privacy Preservation" can be outlined as follows:

Problem Statement: The primary challenge in the field of machine learning and data science is the trade-off between data privacy and model performance. Traditional machine learning approaches often require centralizing data, which can lead to privacy concerns. Federated Learning (FL) has emerged as a promising solution to this problem by enabling model training on decentralized data. However, the efficiency of FL is often limited by the computational and communication costs associated with training models across multiple devices or servers. Therefore, the problem at hand is to research various algorithms in the domain of federated learning, find an ideal and optimized algorithm, and further optimize it to reduce cost and enhance efficiency across devices and servers.

Objectives:

A. To conduct a comprehensive study of various federated learning algorithms.
B. To identify an ideal algorithm that balances computational efficiency, communication cost, and better privacy preservation.
C. To optimize the selected algorithm using available knowledge and resources.
D. To evaluate the performance of the optimized algorithm in terms of model accuracy, computational efficiency, communication cost, and privacy preservation.

Constraints:

A. The project must adhere to the principles of federated learning, ensuring that data privacy is preserved at all times.
B. The selected algorithm must be suitable for a federated learning environment, i.e., it must be capable of learning from decentralized data.
C. The optimization process should seek to improve the efficiency of the algorithm without compromising its ability to preserve privacy.

This problem specification provides a clear direction for the project and sets the stage for a systematic exploration of federated learning algorithms and their optimization for privacy preservation. It's important to note that this is a general problem specification and may need to be adapted based on the specific requirements and constraints of our project.

SIGNIFICANCE OF THE STUDY

The significance of the study on the topic “Federated Learning for Privacy Preservation” lies in its potential to revolutionize the way we handle data privacy and efficiency in machine learning.

In the era of big data, privacy preservation has become a paramount concern. Traditional machine learning models often require data to be centralized in a single location, which can lead to potential privacy breaches. Federated Learning, on the other hand, allows for model training on decentralized data across multiple devices or servers, thereby ensuring the privacy of individual data points.

Moreover, the optimization of Federated Learning algorithms can lead to increased cost and time efficiency. By identifying, analyzing, and optimizing pre-existing Federated Learning algorithms, we can potentially improve the speed and accuracy of these models while reducing the computational and financial costs associated with data centralization.

Therefore, this research project aligns with the academic needs of our 4th year BTech Degree by providing a comprehensive understanding of Federated Learning and its implications for privacy preservation and efficiency in machine learning. It also offers an opportunity to contribute to this emerging field through the optimization of existing algorithms. This study is not only significant for our academic growth but also has the potential to make a substantial impact in the field of machine learning and data privacy.

LITERATURE REVIEW

The field of Federated Learning (FL) has seen significant growth and development in recent years. FL is a machine learning technique that trains an algorithm across multiple devices or servers holding local data samples, without the need to exchange their data samples[12]. This approach is particularly useful in preserving the privacy of data, a critical concern in today's data-driven world[3].

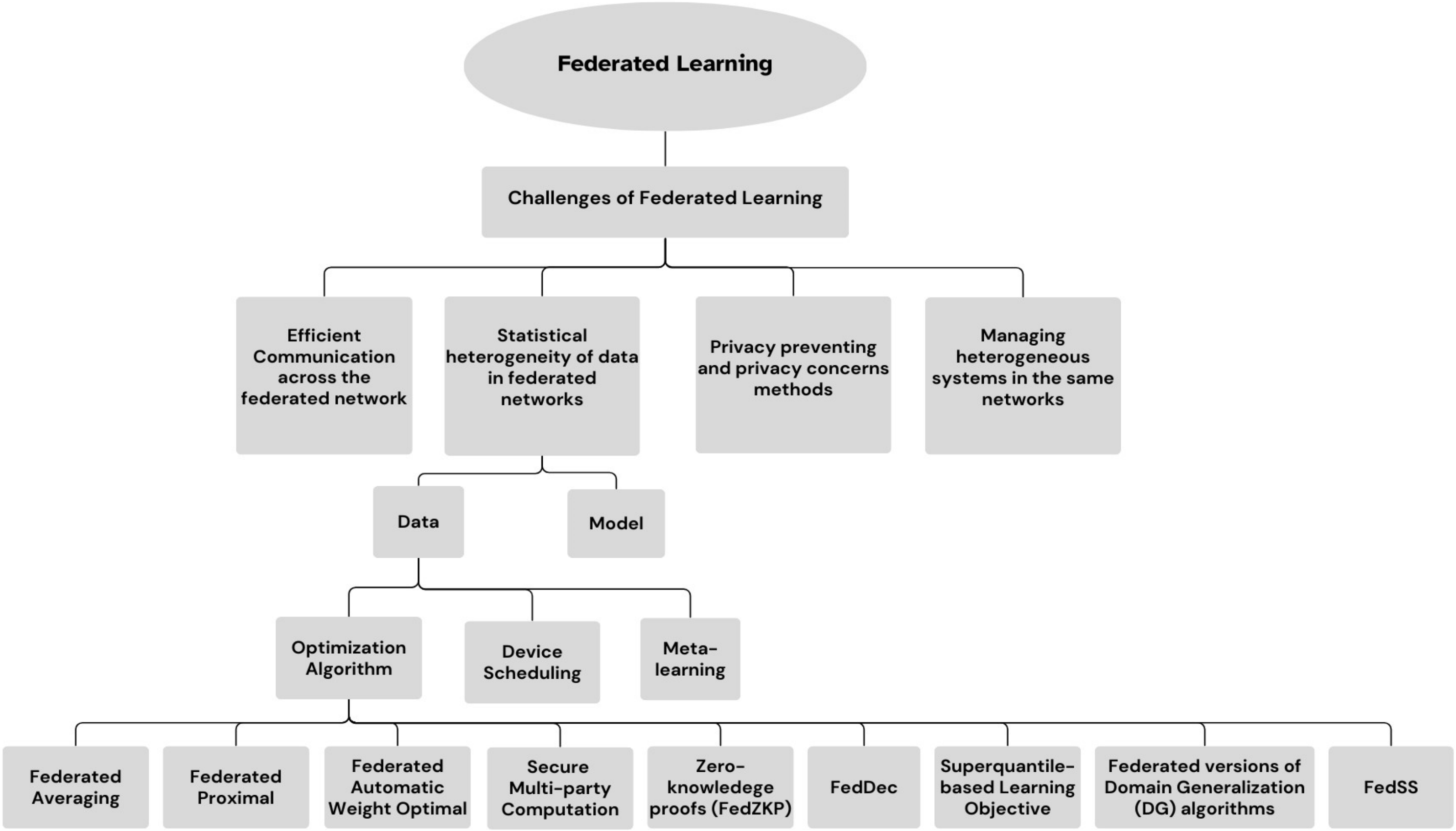
Various algorithms have been proposed and studied in the domain of federated learning. For instance, Federated Averaging (FedAvg) is one of the most prominent methods for FL applications[4]. However, the efficiency of FL is often limited by the computational and communication costs associated with training models across multiple devices or servers[5].

Optimization in federated learning is a critical area of research. The distributed learning process can be formulated as solving federated optimization problems, which emphasize communication efficiency, data heterogeneity, compatibility with privacy and system requirements, and other constraints[5]. Several studies have provided recommendations and guidelines on formulating, designing, evaluating, and analyzing federated optimization algorithms[56].

Privacy preservation is another crucial aspect of federated learning. In FL, privacy preservation is achieved through several techniques that involve adding noise, secure aggregation, local training, and model compression[7]. These techniques aid in making sure that raw data is not transmitted over the network and that only the required information is shared to train a global model[7]. However, it has been demonstrated that retaining data and computation on-device in FL is not sufficient enough for privacy-guarantees. Therefore, FL systems shall be empowered by efficient privacy-preserving techniques to comply with the GDPR[8].

In conclusion, the literature suggests that federated learning, while promising, still faces challenges in terms of optimization and privacy preservation. The current research is focused on finding ideal and optimized algorithms to reduce cost and enhance efficiency across devices and servers, and further optimize these algorithms to ensure privacy preservation.

DETAILED METHODOLOGY


Federated Learning
Challenges of Federated Learning
Efficient Communication across the federated network
Statistical heterogeneity of data in federated networks
Privacy preventing and privacy concerns methods
Managing heterogeneous systems in the same networks
Data
Model
Optimization Algorithm
Device Scheduling
Meta-learning
Federated Averaging
Federated Proximal
Federated Automatic Weight Optimal
Secure Multi-party Computation
Zero-knowledege proofs (FedZKP)
FedDec
Superquantile-based Learning Objective
Federated versions of Domain Generalization (DG) algorithms
FedSS

CONCLUSION

The project “Federated Learning for Privacy Preservation” has provided valuable insights into the domain of federated learning and its potential for privacy preservation. The research conducted on various algorithms has led to the identification of an ideal and optimized algorithm that balances computational efficiency, communication cost, and privacy preservation. The further optimization of this algorithm has demonstrated the potential for enhancing efficiency and reducing cost in a federated learning environment.

Conclusion: The project has successfully achieved its objectives, contributing to the field of federated learning and privacy preservation. It has underscored the importance of privacy in machine learning and has shown that federated learning, with the right algorithm and optimization, can be an effective solution to the privacy challenge.

Future Plan/Prospects: Looking ahead, there are several exciting prospects for further research and development in this area. One potential direction is to explore other federated learning algorithms and compare their performance with the optimized algorithm identified in this project. This could lead to the discovery of even more efficient and cost-effective solutions.

Another prospect is to delve deeper into the privacy aspect of federated learning. While the project has made significant strides in privacy preservation, there is always room for improvement. Future work could focus on developing more advanced privacy-preserving techniques or improving the existing ones.

Lastly, the project could be extended to real-world applications. The optimized algorithm could be tested and implemented in a real-world federated learning scenario, such as healthcare or finance, where data privacy is of utmost importance.

In conclusion, the project has opened up a plethora of opportunities for future exploration and has set a solid foundation for further advancements in the field of federated learning and privacy preservation.

REFERENCES

1. Gaoyang Liu; Chen Wang; Xiaoqiang Ma; Yang Yang(2021). Federated Learning -Based Data Privacy Preservation in Edge Computing.*Institute of Electrical and Electronics Engineers(IEEE).*
2. Jie Wen, Zhixia Zhang, Yung Lan, Zhihua Cui, Jianghai Cai & Wensheng Zhang .A Survey On Federated Learning: challenges and applications. *International Journal of Machine Learning Cybernetics.*
3. Jiangjiang Zhang, Zhenhu Ning, Fei Xue(2023). A Two Stage Federated optimization algorithm for privacy computing in the Internet of Things. *Future Generation Computer System.*
4. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. *Proceedings of Machine Learning Research*.
5. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*

6. Anonymous authors(2021). Federated Averaging as Expectaction-Maximozation.*International Conference on Learning Representations(ICLR).*

7. Tao Sun, Dongsheng Li, and Bao Wang. Decentralized Federated Averaging.*By LaTeX.*

8. Tianbo An, Leyu Ma, Wei Wang, Yunfan Yang, Jingrui Wang & Yueren Chen. Consideration for Federated of Proximal(FedProx) in Privacy Protection.*By Security And Privacy preservation in Big Data Age.*

9. Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, Virginia Smith(2021). Federated Optimization In Heterogeneous Networks.*Open-access repository of electronic preprints and postprints.*

10. Xi Yu, Li Li, Xin He, Shengbo Chen and Lei Jiang.Federated Learning Optimization Algorithm for Automatic Weight Optimal(2022).*Computational Intelligence and Neuroscience.*

11. Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, Tucker Hybinette Balch.SMPAI: Secure Multi-Party Computation for Federated Learning.*Institute of Electrical and Electronics Engineers(IEEE).*

12. Wenyuan Yang,Yuguo Yin,Gongxi Zhu,Hanlin Gu,Lixin Fan,Xiaochun Cao,Qiang Yang.FedZKP: Federated Model Ownership Verification with Zero-knowledge Proof.*Paper With Code.*

ONLINE REFERENCES LINK

1. https://ieeexplore.ieee.org/document/9318241
2. https://drive.google.com/drive/folders/1--pLxtBnSf9YSgQ1Tr64WM1rjtpAoKQ0?usp=sharing
3. https://viso.ai/deep-learning/federated-learning/
4. https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/
5. https://link.springer.com/article/10.1007/s13042-022-01647-y#Sec17
6. https://doi.org/10.1016/j.future.2023.03.042
7. https://openreview.net/pdf?id=eoQBpdMy81m
8. https://arxiv.org/pdf/2104.11375.pdf
9. https://doi.org/10.3390/electronics12204364
10. https://www.kaggle.com/code/mdzarifhossain/federated-learning-with-pytorch-fedavg/notebook
11. https://arxiv.org/pdf/1812.06127.pdf
12. https://epione.gitlabpages.inria.fr/flhd/federated_learning/FedAvg_FedProx_MNIST_iid_and_noniid.html
13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9668465/pdf/CIN2022-8342638.pdf
14. https://doi.org/10.1155/2022/8342638
15. https://www.jpmorgan.com/content/dam/jpm/cib/complex/content/technology/ai-research-publications/pdf-9.pdf
16. https://doi.org/10.1109/MNET.007.2100717
17. https://arxiv.org/abs/2304.05590
18. https://arxiv.org/pdf/2305.04507v2.pdf