International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 2, Issue 6, November-December 2019||

DOI:10.15662/IJRAI.2019.0206002

Federated Learning over Non-IID Edge Devices

Vikash Anand Patel

Ladhidevi Ramdhar Maheshwari Night College of Commerce, Mumbai, Maharashtra, India

ABSTRACT: Federated Learning (FL) enables decentralized model training across edge devices while preserving data privacy. However, the inherent non-Independent and Identically Distributed (non-IID) nature of data on these devices poses significant challenges to model convergence and performance. This paper reviews strategies developed before 2018 to address these challenges, focusing on data-sharing techniques, model aggregation methods, and communication-efficient algorithms. We analyze the effectiveness of these approaches in improving model accuracy and convergence speed under non-IID conditions. Our review highlights the trade-offs between privacy preservation and model performance, providing insights into the evolution of FL methodologies.

KEYWORDS: Federated Learning, Non-IID Data, Edge Devices, Model Aggregation, Data Privacy, Communication Efficiency

I. INTRODUCTION

Federated Learning (FL) is a decentralized machine learning paradigm that allows multiple edge devices to collaboratively train a global model without sharing their local data. This approach is particularly beneficial in scenarios where data privacy is paramount, such as in healthcare and finance. However, one of the significant challenges in FL is the non-IID nature of data across devices. In non-IID settings, data distributions vary significantly between devices, leading to issues like model divergence and slow convergence. Traditional FL algorithms, such as Federated Averaging (FedAvg), assume IID data and may not perform well under non-IID conditions. Therefore, addressing the non-IID challenge is crucial for the practical deployment of FL in real-world applications.

II. LITERATURE REVIEW

Before 2018, several approaches were proposed to mitigate the impact of non-IID data in FL. Zhao et al. (2018) demonstrated that the accuracy of FL models could decrease by up to 55% when trained on highly skewed non-IID data. They attributed this performance degradation to weight divergence among local models and proposed a strategy of sharing a small subset of globally representative data to improve model accuracy. Similarly, McMahan et al. (2016) introduced FedAvg, a communication-efficient FL algorithm, and showed its robustness to unbalanced and non-IID data distributions. However, they noted that the algorithm's performance could still be affected by significant data heterogeneity. These studies laid the groundwork for understanding and addressing the challenges posed by non-IID data in FL.

III. RESEARCH METHODOLOGY

This study employs a qualitative research methodology, conducting a comprehensive review of existing literature on FL techniques addressing non-IID data. The review focuses on studies published before 2018, analyzing their proposed solutions, experimental setups, and outcomes. Key aspects such as data distribution strategies, model aggregation methods, and communication protocols are examined to assess their effectiveness in improving FL performance under non-IID conditions. The findings are synthesized to provide insights into the evolution of FL methodologies and to identify gaps for future research.

IV. ADVANTAGES

- **Data Privacy Preservation**: FL allows model training without sharing raw data, ensuring privacy compliance.
- **Scalability**: FL can scale across a large number of edge devices, leveraging distributed computing resources.
- **Reduced Communication Overhead**: By sharing model updates instead of data, FL reduces the amount of data transmitted over the network.

| ISSN: 2455-1864 | www.ijrai.com | editor@ijrai.com | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 2, Issue 6, November-December 2019||

DOI:10.15662/IJRAI.2019.0206002

V. DISADVANTAGES

- **Non-IID Data Challenges**: Data heterogeneity across devices can lead to model divergence and poor generalization.
- **Communication Bottlenecks**: Frequent communication between devices and the central server can lead to latency and bandwidth issues.
- **Device Availability**: The intermittent availability of edge devices can affect the consistency and reliability of model training.

VI. RESULTS AND DISCUSSION

Studies have shown that sharing a small subset of globally representative data can significantly improve model accuracy in non-IID settings. For instance, Zhao et al. (2018) reported a 30% increase in accuracy on the CIFAR-10 dataset by sharing just 5% of the data. However, this approach may compromise data privacy, as it involves centralizing a portion of the data. Alternatively, communication-efficient algorithms like FedAvg have been proposed to reduce communication overhead, though their performance can still be affected by data heterogeneity. These findings underscore the need for a balance between privacy preservation and model performance in FL systems.

VII. CONCLUSION

Addressing the non-IID data challenge is crucial for the effective deployment of Federated Learning in real-world applications. While various strategies have been proposed, each comes with its trade-offs between model performance and data privacy. Future research should focus on developing novel algorithms that can effectively handle non-IID data without compromising privacy, scalability, or communication efficiency.

VIII. FUTURE WORK

Future research directions include:

- **Development of Privacy-Preserving Algorithms**: Designing algorithms that can handle non-IID data without sharing sensitive information.
- **Improved Model Aggregation Techniques**: Creating aggregation methods that are robust to data heterogeneity.
- **Adaptive Communication Protocols**: Developing protocols that can dynamically adjust to varying network conditions and device availability.

REFERENCES

1. McMahan, H. B., et al. (2016). Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629.
2. Zhao, Y., et al. (2018). Federated learning with non-IID data. arXiv preprint arXiv:1806.00582.