CS4681 - Advanced Machine Learning

Progress Report

Communication-Efficient Learning of Deep Networks from Decentralized Data

Index No : 210613U

Name : S.M.C.R.Siriwardhana

1. Project Overview

This project aims to enhance the Federated Learning Framework , as implemented in the Communication-Efficient Learning of Deep Networks from Decentralized Data. research paper[1]. The primary goal is to address the algorithm's key limitations: degraded performance on heterogeneous (non-IID) data and high communication overhead. This will be achieved through two primary enhancements: first, by replacing traditional Multi-Layer Perceptron (MLP) layers with the novel Kolmogorov-Arnold Networks (KANs) to improve model performance and interpretability; and second, by integrating model compression techniques like quantization and pruning to reduce the size of model updates and speed up training. The project will culminate in a research paper demonstrating measurable performance gains over the FedAvg baseline on realistic federated datasets.

2. Literature Review

2.1. The Federated Learning Imperative

Federated learning (FL) is a pivotal distributed machine learning concept that enables collaborative model training on decentralized data while preserving privacy. The central server coordinates a "loose federation" of clients (e.g., mobile devices), each with a private local dataset that is never uploaded to the server. This approach is a direct application of the principle of data minimization, decoupling model training from the need for direct access to raw data and significantly reducing the attack surface for privacy and security risks.

At the heart of this paradigm lies the Federated Averaging (FedAvg) algorithm, a practical and

widely adopted method for deep network training in a decentralized setting. FedAvg's core innovation is its focus on mitigating communication costs, which are the principal constraint in federated optimization. The algorithm operates by allowing clients to perform multiple local updates using their private datasets before transmitting a single, aggregated update to a central server . The server then averages these updates to refine a global model. This approach drastically reduces the number of communication rounds required for convergence by a factor of 10-100x compared to traditional synchronized stochastic gradient descent (SGD) ..[11][12][13]

2.2. Challenges of Federated Learning

FedAvg's design is a response to unique challenges in federated optimization:

- **Communication Efficiency:** This is the most significant bottleneck in the federated setting. Mobile devices are often offline, on slow, or on expensive connections, making frequent, large-scale data transfers untenable. The underlying assumption of FedAvg is that local computation is "essentially free" compared to the cost of communication, so the algorithm deliberately trades additional local computation for a reduction in costly communication rounds[1].
- **Statistical Heterogeneity (Non-IID Data):** In a federated network, data on any given client is a function of a particular user's behavior, meaning the local dataset is rarely a representative sample of the overall population. This violates the independent and identically distributed (IID) data assumption made by most distributed optimization algorithms. The downstream effects are severe, including model bias towards clients with more prominent data, slower convergence as the global model struggles to reconcile divergent local updates, and reduced accuracy on underrepresented data subsets[3]. The original FedAvg paper[1] acknowledged that this is a "pathological" problem, particularly when averaging models trained on entirely distinct data distributions.
- **Systems Heterogeneity** : Beyond data distribution, the physical characteristics of devices themselves vary widely. Differences in hardware (CPU, memory), network connectivity, and power levels can lead to some clients being "stragglers" or dropping out entirely during training . An effective FL system must be robust to this variability and capable of handling a low amount of client participation per round .

2.3. Enhancement Methodologies

This outlines a methodology to enhance the given framework by addressing its primary weaknesses: performance on heterogeneous data and communication overhead. The proposed enhancements focus on two key areas:

1. **Novel Architectural Paradigms:** The implementation of Kolmogorov-Arnold Networks (KANs) as a replacement for traditional Multi-Layer Perceptrons (MLPs). This is a cutting-edge approach that promises superior functional approximation and interpretability.[4][5]
2. **Communication Efficiency:** The integration of model compression techniques, specifically quantization and pruning, to reduce the size of model updates and alleviate the communication bottleneck.[6]

3. Methodology Outline

The project will implement two distinct, complementary enhancement strategies to the FedAvg baseline.

3.1. Architectural Enhancement: Integrating Kolmogorov-Arnold Networks (KANs)

Kolmogorov-Arnold Networks (KANs) represent a significant architectural departure from traditional neural networks, offering a fresh perspective on function approximation.[5] They replace linear weights and fixed activation functions of conventional MLPs with learnable, univariate functions on the network's edges, often parameterized as B-splines.

Integration Method:
The integration of KANs, termed Fed-KAN, involves replacing the MLP layers in a conventional model with KAN layers. For image classification, a hybrid model would use conventional convolutional layers for feature extraction and a KAN-based component for the classification head.[10]

Benefits and Trade-offs:

- **Superior Performance on Non-IID Data:** Empirical studies provide compelling evidence that Fed-KAN models can outperform their conventional Fed-MLP counterparts, particularly in heterogeneous, non-IID data scenarios. The superior functional approximation capabilities of KANs allow each local client model to better adapt to its unique data distribution.[4]
- **Enhanced Interpretability:** KANs possess an unparalleled level of interpretability, as their learned B-spline functions can be simplified to reveal underlying symbolic formulas.[4]
- **Computational Bottleneck :** A significant drawback is the computational overhead of KANs. Their spline-based functions are more complex to compute than simple matrix multiplications, and they are not currently optimized for GPU parallelization. This

contradicts FedAvg's core assumption that local computation is "essentially free".[12]

- **Less Training Rounds:** The direct communication cost per round (model size) for KANs are higher than that of the MLP but KAN would converge for an expected performance level faster than a traditional MLP. Therefore the Overall Communication cost would reduce by considerable amounts.[12]

3.2. Communication Enhancement: Integrating Model Compression

The communication overhead of federated learning, where model updates can be large and network connections slow, is a major bottleneck. Model compression techniques offer a direct solution by reducing the size of the messages exchanged between clients and the server.

Model Compression Techniques:

- **Quantization:** This is a process that reduces the numerical precision of model updates from high-precision formats (e.g., 32-bit floating-point) to lower-bit representations. This can drastically cut down the size of the messages.[6][13]
- **Pruning:** Pruning creates sparse models by removing parameters (weights) that are deemed less important. This reduces the total number of values that need to be transmitted, thereby lowering communication overhead and potentially computation costs[7]. A sophisticated example is the SpaFL framework, which tackles this by introducing trainable thresholds for each neuron or filter. Instead of communicating model parameters, SpaFL only communicates these thresholds, which are orders of magnitude smaller.

Technique	**Primary Mechanism**	**Key Benefit**	**Main Trade-off/Challenge**
Quantization	Reduces the bit precision of model updates.	Drastically cuts communication overhead.	Accuracy degradation due to precision loss.
Pruning	Removes unimportant parameters to create sparse models.	Reduces model size, lowering both communication and computation.	Can compromise model performance on non-IID data.
SpaFL	Only communicates trainable thresholds	Achieves massive communication	More complex algorithm that

	instead of parameters.	savings (e.g., 0.17% of FedAvg) and improves accuracy.	decouples pruning from model parameters.

Table 1: Key Communication Efficiency Techniques

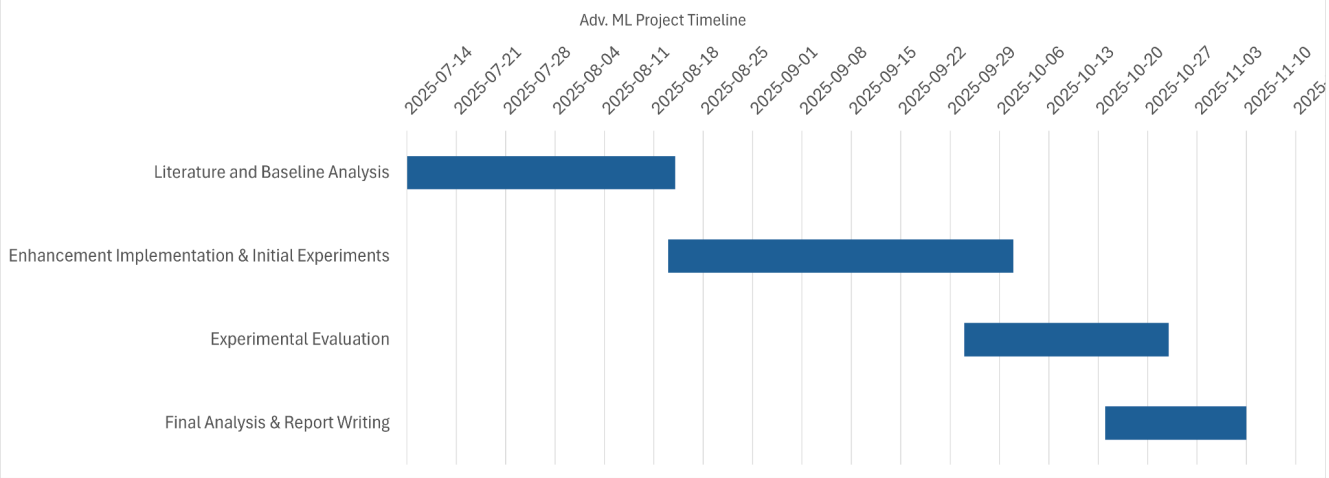
4. Project Plan and Timeline

The project will follow a structured 12-week timeline, divided into four distinct phases.

Week(s)	Phase	Description
1-4	**Phase 1: Literature and Baseline Analysis**	Comprehensive literature review. Set up LEAF framework and reproduce FedAvg baseline on FEMNIST and Shakespeare datasets.
5-9	**Phase 2: Enhancement Implementation & Initial Experiments**	Implement KAN architecture and model compression techniques (quantization/pruning) within the LEAF framework. Conduct initial experiments to verify functionality.
10-11	**Phase 3: Experimental Evaluation**	Systematically compare enhanced models against defined baselines. Perform hyperparameter tuning and meticulously collect quantitative data on all metrics (accuracy, communication rounds, etc.).

12	**Phase 4: Final Analysis & Report Writing**	Synthesize experimental results. Generate graphs of performance gains. Author the final research paper, including a detailed discussion of the findings and trade-offs.

Table 2 : Project Plan



5. Conclusions and Future Work

This project provides a clear and actionable plan for enhancing the foundational FedAvg algorithm by addressing its two most significant limitations: performance on heterogeneous data and high communication overhead. By adopting a novel architectural paradigm (KANs) and integrating model compression techniques (quantization and pruning), the proposed methodology offers a powerful and multi-faceted solution.

This project will not only validate the effectiveness of these enhancements but will also contribute to a deeper understanding of the trade-offs involved. The findings will provide empirical evidence on how a practitioner must choose between the potential for groundbreaking performance and interpretability of KANs versus the proven communication efficiency of model compression techniques. Future work could explore combining these two approaches, such as integrating quantized or pruned KANs into a federated framework, to achieve a system that is

simultaneously high-performing, interpretable, and communication-efficient.

References

1. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. a. Y. (2016, February 17). Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv.org. https://arxiv.org/abs/1602.05629
2. Strategy in Federated Learning by William Lindskog, https://pub.towardsai.net/1-strategy-in-federated-learning-e0e518ada44a
3. Non-IID data in Federated Learning: A Systematic Review with Taxonomy, Metrics, Methods, Frameworks and Future Directions - arXiv, https://arxiv.org/html/2411.12377v1
4. KAN: Kolmogorov–Arnold Networks - OpenReview, https://openreview.net/forum?id=Ozo7qJ5vZi
5. Kolmogorov-Arnold Networks (KANs): A Guide With Implementation | DataCamp, https://www.datacamp.com/tutorial/kolmogorov-arnold-networks
6. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization, https://proceedings.mlr.press/v108/reisizadeh20a.html
7. Model Pruning Enables Efficient Federated Learning on Edge Devices - Shiqiang Wang, https://shiqiang.wang/papers/YJ_SpicyFL2020.pdf
8. A Beginner-friendly Introduction to Kolmogorov Arnold Networks (KAN), https://www.dailydoseofds.com/a-beginner-friendly-introduction-to-kolmogorov-arnold-networks-kan/
9. Understanding Kolmogorov Arnold Networks (KAN) | TDS Archive - Medium, https://medium.com/data-science/understanding-kolmogorov-arnold-networks-kan-e317b1b4d075
10. Convolutional Kolmogorov-Arnold Networks (Convolutional KANs): An Innovative Alternative to the Standard Convolutional Neural Networks (CNNs) - MarkTechPost, https://www.marktechpost.com/2024/06/24/convolutional-kolmogorov-arnold-networks-convolutional-kans-an-innovative-alternative-to-the-standard-convolutional-neural-networks-cnns/
11. Fed-KAN: Federated Learning with Kolmogorov-Arnold Networks for Traffic Prediction, https://arxiv.org/html/2503.00154v1
12. Kolmogorov-Arnold Networks (KAN): Alternative to Multi-Layer Perceptron? - DigitalOcean, https://www.digitalocean.com/community/tutorials/kolmogorov-arnold-networks-kan-revolutionizing-deep-learning
13. Quantization For Federated Learning - Meegle, https://www.meegle.com/en_us/topics/quantization/quantization-for-federated-learning
14. Fed-KAN: Federated Learning with Kolmogorov-Arnold Networks, https://www.researchgate.net/publication/389547173_Fed-KAN_Federated_Learning_with_Kolmogorov-Arnold_Networks_for_Traffic_Prediction