Tian Zhao

PROPOSAL FOR CAPSTONE PROJECT

Project Title: Federated Learning for Image Classification with a Focus on Data Distribution and Performance Factors

Chris Harry Patrick - MS in CS (Regular Track)

Capstone Supervisor: Dr. Tian Zhao

Introduction:

Federated Learning (FL) has gained significant attention as a decentralized approach to machine learning, where the model is trained on edge devices, ensuring data privacy by keeping data local. A key application of FL is image classification, which faces challenges when the data distribution across devices is non-IID (Independent and Identically Distributed). Additionally, the number of participating clients and the frequency of model aggregation significantly impact FL's performance.

While existing research has addressed federated learning with non-IID data, there remains a gap in understanding how **dynamic, time-varying data distributions**—such as those resulting from device mobility or environmental changes—impact model accuracy and efficiency. Furthermore, the effect of **adaptive client participation**—considering device heterogeneity (processing power, network conditions)—and **aggregation frequency** in real-world federated learning settings is underexplored.

In this project, I aim to investigate these factors to improve the efficiency and accuracy of federated image classification models in real-world, resource-constrained environments.

Research Questions:

1. **How does dynamic data distribution affect federated learning for image classification across devices?**
    - Existing research has primarily focused on static non-IID data. I plan to explore how **time-varying, dynamic data distributions** (e.g., due to device mobility, environmental changes, or user activity) influence model accuracy and convergence rates in federated learning setups.
2. **How does adaptive client participation, based on device heterogeneity, and model aggregation frequency impact federated image classification performance?**
    - While previous work has varied the number of clients, it has not adequately considered **client heterogeneity**, such as varying processing power or network conditions. This project will test how **adaptive client participation** and **dynamic aggregation frequencies** affect model performance, communication cost, and convergence in federated learning

Proposed Methodology:

1. **Dataset Selection**:
    - I will use **MNIST** and **CIFAR-10**, standard image classification datasets, to simulate dynamic data distributions across different clients. These datasets are widely used in federated learning studies and will allow for easy benchmarking and comparison of performance.
2. **Experimentation with Parameters**:
    - **Dynamic Data Distribution**: I will simulate **time-varying data distributions** by modifying the data at regular intervals to mimic real-world scenarios (e.g., environmental changes, device mobility).
    - **Adaptive Client Participation**: I will vary the number of participating clients based on client capabilities (e.g., device processing power and network conditions) to study its effect on federated learning performance.
    - **Aggregation Frequency**: I will experiment with different aggregation frequencies (e.g., local updates every few rounds vs. frequent aggregations) to explore trade-offs between communication efficiency and model accuracy.
3. **Performance Metrics**:
    - **Accuracy and Loss**: Evaluate model accuracy and loss over different rounds of training.
    - **Communication Cost**: Analyze the trade-off between communication efficiency (frequency of aggregation and updates) and performance.
    - **Convergence Rates**: Measure how quickly the model converges under different settings.
4. **Model Training**:
    - I will employ simulation frameworks to design and run experiments.

Expected Outcomes:

- **Insights into Dynamic Data Distributions**: A better understanding of how time-varying data distributions impact federated learning models for image classification and how to handle such data more effectively.
- **Recommendations for Adaptive Client Participation**: A clearer view of how client heterogeneity affects model performance and communication efficiency, with practical guidelines for adaptive participation strategies.
- **Comparative Analysis of Aggregation Frequency**: A practical evaluation of how different aggregation frequencies influence the trade-off between communication efficiency and model accuracy.
- **Real-World Applicability**: Insights that can be applied to federated learning in **resource-constrained environments** such as IoT devices, where both data and client resources are limited.

Significance:

This project aims to fill critical gaps in federated learning research by focusing on **dynamic, time-varying data distributions**, **adaptive client participation**, and

aggregation frequency. The results will offer a comprehensive understanding of the factors that influence federated learning performance in practical scenarios, such as mobile devices or IoT networks. The insights gained can directly inform the design of more efficient and scalable federated learning systems, particularly in resource-constrained environments.

Frequency of Meetings:

I plan to meet with my supervisor bi-weekly to discuss progress, receive feedback, and address challenges encountered during the project. Additional meetings can be scheduled as needed.

Evaluation Criteria:

The evaluation of my project will be based on:

- **Quality of Experimentation**: Depth of analysis and the rigor of experimental design.
- **Clarity and Accuracy of Findings**: How clearly the results are presented and the validity of the conclusions.
- **Completeness of the Final Report**: Including data visualizations, analysis, and insights.
- **Adherence to Timeline and Milestones**: Meeting key deadlines and demonstrating progress in line with project goals.

Novel Contribution and Differences from Existing Research:

1. **Dynamic Data Distribution**:
    - **Existing Research**: Federated learning with non-IID data is a well-explored topic, focusing on static data distributions.
    - **Novelty**: My project will specifically explore **dynamic, time-varying data distributions**, simulating real-world scenarios where data across devices is constantly changing, affecting model training.
2. **Adaptive Client Participation**:
    - **Existing Research**: Many studies focus on the number of clients but do not consider the heterogeneity of clients in terms of processing power, network conditions, and device capabilities.
    - **Novelty**: This project will introduce the concept of **adaptive client participation** based on device capabilities, offering a new angle to understand how client resources affect federated learning performance.
3. **Aggregation Frequency**:
    - **Existing Research**: While aggregation frequency has been discussed, the focus has mostly been on the impact of client numbers, with less attention paid to how **aggregation frequency** and **client participation** interact in federated learning.
    - **Novelty**: This project will specifically test and empirically evaluate the impact of **varying aggregation frequencies** combined with dynamic client participation, providing practical recommendations for real-world federated learning systems.

References:

1. McMahan, B., et al. "Communication-efficient learning of deep networks from decentralized data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, 2017.
2. Kairouz, P., et al. "Advances and open problems in federated learning." arXiv preprint arXiv:1912.04977, 2019.
3. Li, T., et al. "Federated optimization in heterogeneous networks." Proceedings of Machine Learning and Systems 2 (2020): 429-450.
4. Zhao, Y., et al. "Federated learning with non-IID data." arXiv preprint arXiv:1806.00582, 2018.
5. Bonawitz, K., et al. "Towards federated learning at scale: System design." Proceedings of the 2nd SysML Conference, 2019.