Improving the Accuracy of Federated Learning Simulations

Using Traces from Real-world Deployments to Enhance the Realism of Simulation Environments

CSE3000 Research project, 2024
Author: **Alexander Nygård**
Responsible professor: **Jérémie Decouchant**
Supervisor: **Bart Cox**

Introduction

Federated Learning (FL) - An ML technique enabling decentralized devices to collaboratively learn a shared model without sharing local data [1].

- **Challenges:** High development costs due to deployment requirements and privacy considerations [2], making simulation a valuable tool.

FL Simulation - Allows development and testing of algorithms and topologies in a controlled environment [3].

- **Current Research Gap:** Previous studies have not sufficiently focused on how representative simulations are of real-world conditions.

Contribution - Quantitatively measure how incorporating attributes from pseudo-real deployments can improve the accuracy of simulations.

- **Impacting Factors:** Identify and understand the factors that most significantly affect simulation realism.

Methodology

Goal: Capture information on non-IID attributes in a pseudo-real FL deployment, then recreate individual attributes in simulations.
Outcome: Understanding of which characteristics of the deployment are most useful to harness in simulators to make them more representative of the real world.

Experiment configurations:

- 1 deployment, *fully non-IID*
- 1 'real' simulation, *fully non-IID*
- 4 varying simulations, *one non-IID attribute*
- 1 'blind' simulation, *fully IID*

Configurations are visualized in Table 1.

Non-IID variables, generated and collected in the deployment (Fig 1):

- Batch sizes
- Local epochs
- Data volume
- Data labels

Evaluation metrics

- Centralized accuracy over rounds
- Centralized loss over rounds
- MSE of accuracy training progressions

All configurations use 100 communication rounds, 20 clients, using a 50% sampling rate, and run 5 times for averaging.

A simple CNN is used with the FedAvg aggregation algorithm.

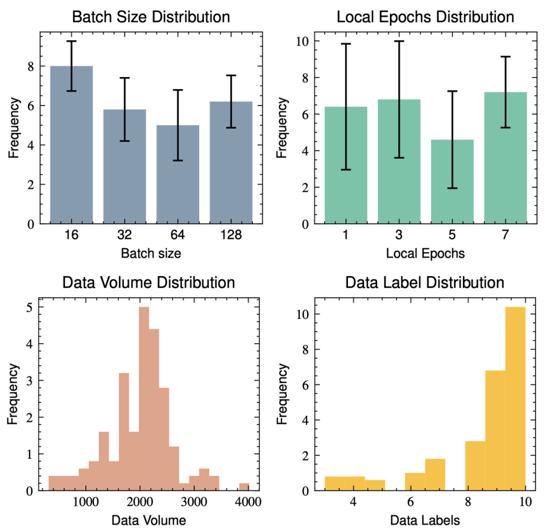

Fig 1. Non-IID uniform client (top) and Dirichlet data (bottom) attribute heterogeneity in a 20-client setting

Experiment	BS	LE	DV	DL
Deployment	Non-IID	Non-IID	Non-IID	Non-IID
Sim. real	Non-IID	Non-IID	Non-IID	Non-IID
Sim. with BS	Non-IID	IID	IID	IID
Sim. with LE	IID	Non-IID	IID	IID
Sim. with DV	IID	IID	Non-IID	IID
Sim. with DL	IID	IID	IID	Non-IID
Sim. blind	IID	IID	IID	IID

Table 1. Comparison of the distributions used for varying experiment configurations. BS: Batch size, LE: Local epochs, DV: Data volume, DL: Data labels.

Results

Key insight 1:
There is a fundamental difference between the deployment and the identical 'real' simulation

Difference of 3.35% accuracy and 0.08 loss (Table 2), and MSE of 21.1 (Fig. 3). Training diverges after ~20th round (Fig. 2 & 4).

Further investigation needed to understand the discrepancy.

Configuration	Accuracy	Loss
Deployment	53.77% ± 2.43%	1.28 ± 0.06
Sim. Real	57.12% ± 2.48%	1.20 ± 0.07
Sim. Batch Size	59.09% ± 1.58%	1.15 ± 0.04
Sim. Local Epochs	59.17% ± 2.23%	1.15 ± 0.06
Sim. Data Volume	60.26% ± 0.96%	1.12 ± 0.03
Sim. Data Labels	57.12% ± 1.27%	1.21 ± 0.03
Sim. Blind	59.92% ± 1.42%	1.13 ± 0.05

Table 2. Best model performance metrics for each configuration, averaged over 5 runs.

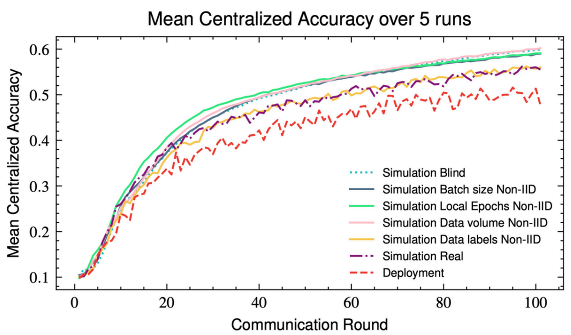

Fig 2. Centralized loss progression over communication rounds for all configurations. Averaged over 5 runs.

Key insight 2:
Including non-IID data labels has most significant impact on recreating deployment results

Smallest difference in performance vs. deployment, Δ3.35% accuracy and Δ0.08 loss. Significant improvement over 'blind' IID simulation, with Δ6.15% and Δ0.15, respectively (Table 2).

The simulation with only non-IID data labels is almost identical to the 'real' simulation (equal accuracy, loss Δ of 0.01, and MSE of 1.6), and has the closest MSE to the deployment with 20.0. This indicates the strong impact of non-IID data labels on simulation realism.

Key insight 3:
Including non-IID batch sizes, local epochs, and data volumes has an insignificant effect on simulation realism

Similar performance and convergence outcomes to the 'blind' fully IID simulation.

All within <1.2% accuracy and 0.03 loss difference, and a max MSE of 3.0 between them.

This indicates that using traces of non-IID batch sizes, local epochs, and data volumes has an insignificant effect on recreating the deployment results, because no improvement is made over the naive IID simulation.

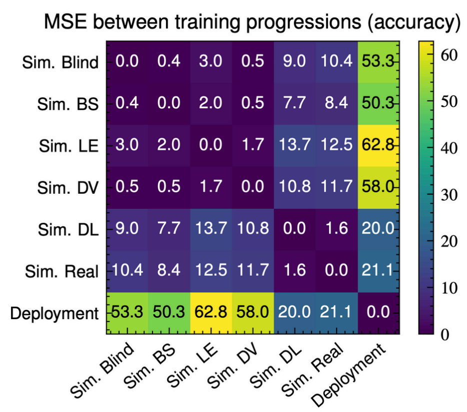

Fig 3. Mean squared error (MSE) between accuracy progressions for each experiment configuration. BS: Batch size, LE: Local epochs, DV: Data volume, DL: Data labels. (Accuracies have been scaled from [0,1] to [0,100] during calculation, for readability.)

Results (cont.)

In summary, **key insight 1** reveals the discrepancies between identical simulations and deployments.

Moreover, **key insights 2 and 3** reveal the strong impact of non-IID data labels, and the insignificant effect of non-IID batch sizes, local epochs, and data volumes.

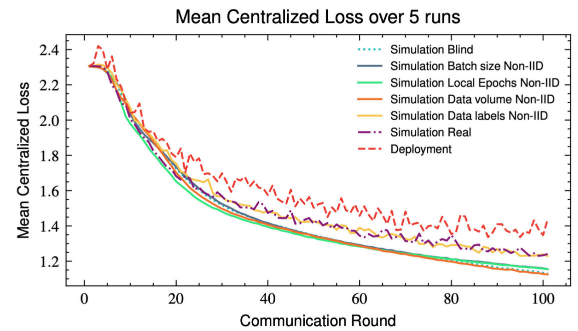

Fig 4. Centralized loss progression over communication rounds for all configurations. Averaged over 5 runs.

Therefore, we note the **sole importance of non-IID data labels** in improving the realism of FL simulations, among the tested attributes.

Discussion

Limitations

- **Pseudo-realness**: The study uses a pseudo-real deployment due to difficulties in accessing real-world client data, offering controlled but less realistic scenarios.
- **Generalisability**: The simplistic FL setup with few clients and narrow configurations limits the study's applicability and generalizability.

Future Work

- **Simulation-deployment discrepancies**: Investigate differences between simulations and real deployments with identical setups.
- **Data collection**: Develop privacy-preserving methods to gather data from real-world FL deployments to enhance realism.

Conclusion

Key Findings: Discrepancies exist between non-IID deployments and identical simulations. Non-IID data labels are most impactful in recreating deployment outcomes in simulations, and including non-IID batch sizes, local epochs, and data volumes appears insignificant.
Limitations: The study's small scale and pseudo-real deployment limit applicability to larger, real-world FL systems.
Practical Insights: Insights can enhance FL system development by using deployment data to create realistic benchmarks and improve data collection methods while ensuring user privacy in both academic and commercial applications.

References

[1] McMahan, Brendan, Eider, Moore, Daniel, Ramage, Seth, Hampson, Blaise Aguera y, Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, 2017.
[2] Wen, Jie, Zhixia, Zhang, Yang, Lan, Zhihua, Cui, Jianghui, Cai, Wensheng, Zhang. "A survey on federated learning: challenges and applications". International Journal of Machine Learning and Cybernetics 14. 2(2023): 513–535.
[3] Lorenzo Sani, , Pedro Porto Buarque de Gusmão, Alex Iacob, Wanru Zhao, Xinchi Qiu, Yan Gao, Javier Fernandez-Marques, Nicholas Donald Lane. "Pollen: High-throughput Simulation of Federated Learning via Resource-Aware Client Placement." (2024).