Enhancing Federated Models with Synthetic Data

Author: Pil Kyu Cho
Supervisor: Swier Garst
Responsible Professor: David Tax

1) Introduction to GenFL

- **Federated Learning (FL)** - distributed machine learning approach that allows the global model to train on each client locally without compromising the data privacy of the clients' local datasets.
- **Independent and Identically Distributed (IID) data** - samples are independently drawn from a fixed distribution and each client dataset is assumed to have the same distribution.
- Most real-world data is non-IID **[1]**, which affects model's performance in FL due to clients having imbalanced datasets.
- **GenFL** - newly proposed generative approach to FL where we pre-train the downstream model with synthetic data created by a generative model.

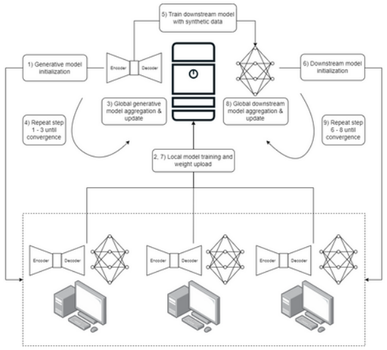

Fig 1: Step-by-step the process of GenFL

2) Research Question

- What is the performance difference for downstream models that were pre-trained on synthetic data in the central server compared to those that did not?
- What is the performance impact on GenFL as data imbalance increases?

3) Research Methodology

- Generative model - **Conditional Variational Autoencoder**
- Datasets - **MNIST**, **FMNIST**
- Data imbalance method - **Dirichlet distribution**
- Generative model metric - image **Classification Accuracy Score [3]**
- FL aggregation algorithm - **FedAVG**
- Downstream model - **ExquisiteNetV1**
- Downstream model metric - **Weighted average of clients' accuracies**

4) Experiment Results

Dataset distribution	MNIST	FMNIST
non-federated IID (baseline)	89.03 ±0.33	65.27 ±0.52
federated IID (baseline)	87.09 ±0.17	62.11 ±0.27
federated non-IID ($\alpha = 100$)	87.48 ±0.52	63.41 ±0.13
federated non-IID ($\alpha = 10$)	86.77 ±0.32	63.83 ±0.4
federated non-IID ($\alpha = 1$)	86.87 ±0.56	63.9 ±0.28
federated non-IID ($\alpha = 0.1$)	84.38 ±0.3	62.85 ±0.09
federated non-IID ($\alpha = 0.01$)	60.65 ±1.16	48.71 ±0.47

Table 1: CAS of CVAE after 20 communication rounds

- As the data imbalance increases, the average CAS of the CVAE decreases.
- Decrease in CAS not noticeable, except for the most extreme data imbalance when the Dirichlet α = 0.01.

Dataset distribution	MNIST		FashionMNIST	
	Without pre-training	With pre-training	Without pre-training	With pre-training
federated IID	87.08 ±1.03	93.92 ±0.28	71.35 ±1.02	78.21 ±0.24
federated non-IID ($\alpha = 100$)	83.78 ±5.32	93.80 ±0.33	70.26 ±1.93	78.32 ±0.13
federated non-IID ($\alpha = 10$)	81.70 ±8.91	93.85 ±0.16	70.23 ±2.55	78.0 ±0.21
federated non-IID ($\alpha = 1$)	85.87 ±2.87	93.59 ±0.26	69.48 ±2.64	77.7 ±0.21
federated non-IID ($\alpha = 0.1$)	41.35 ±15.28	88.24 ±2.43	56.94 ±4.32	68.96 ±3.42
federated non-IID ($\alpha = 0.01$)	22.52 ±5.24	57.45 ±6.67	30.9 ±4.9	48.85 ±8.39

Table 2: Accuracy of ExquisiteNetV1 after 10 communication rounds

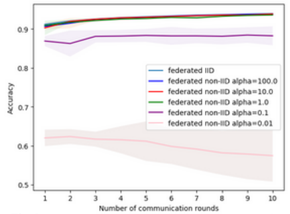

Fig 2: pre-trained ExquisiteNetV1 for MNIST

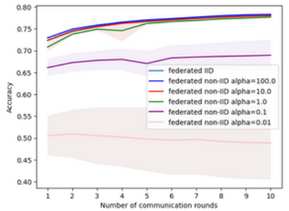

Fig 3: pre-trained ExquisiteNetV1 for FMNIST

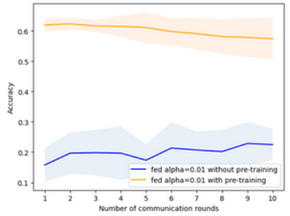

Fig 4: with, without pre-train ExquisiteNetV1 for MNIST

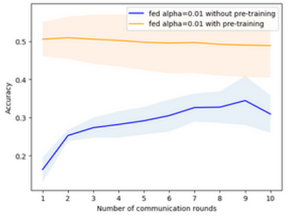

Fig 5: with, without pre-train ExquisiteNetV1 for FMNIST

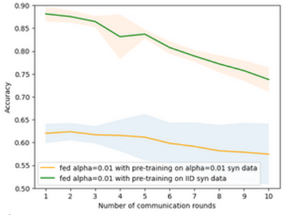

Fig 6: diff syn data pre-train ExquisiteNetV1 for MNIST

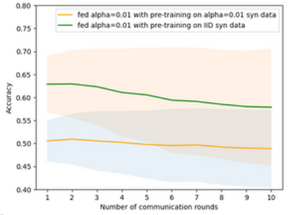

Fig 7: diff syn data pre-train ExquisiteNetV1 for FMNIST

- **Table 2** shows that pre-training with synthetic data before the FL process has a noticeable impact compared to results that didn't.
- As data imbalance increases, so does standard deviation of accuracy, indicating increasing instability with the model.
- **Fig 2**, **3** show when Dirichlet α = 0.01 with pre-training, accuracy average decreases and standard deviation increases over time.
- **Fig 4, 5** show when Dirichlet α = 0.01 without pre-training, accuracy average does not decrease but standard deviation still increases.
- Attempted to mitigate this abnormal behavior when Dirichlet α = 0.01 by pre-training with better quality synthetic data (IID). However, **fig 6**, **7** show same abnormal behavior more clearly.
- One explanation for this abnormal behavior is due to the extreme class and quantity imbalance when Dirichlet α = 0.01. As a result, local training results in extremely biased local weights and simple FedAVG without weight adjustments for imbalance mitigation degrades the model over time.

5) Conclusion of GenFL

- GenFL can be an effective approach to bridging the performance gap that occurs when data is imbalanced amongst the clients.
- As data imbalance increases, GenFL becomes less effective and does not prevent increasing model instability.
- Even with access to high quality data for pre-training, can not mitigate the performance degradation when data imbalance is extreme (i.e. Dirichlet α = 0.01).
- When working with extreme data imbalances, recommended to only initialize the downstream model with pre-trained weights and no local training on clients, as this can degrade the performance of the downstream model.

6) Limitations and Future Work

- Future works can consider more advanced datasets (e.g. CIFAR10) and different generative models (e.g. GAN). **[2]** suggests we can expect similar results as this research.
- Explore other types of non-IID in datasets, such as feature imbalance, that are prevalent in the real world **[1]**.
- Using generative models raises privacy issues of extracting real data used in training. Introducing differential privacy, such as adding noise during generative process, can mitigate this but with some trade off with performance **[4]**.
- Consider incorporating weight adjustment methods to mitigate class imbalances when non-IID is high.

7) References

[1] Q. Li, Y. Diao, Q. Chen, and B. He, Federated learning on non-iid data silos: An experimental study, 2021. DOI:10.48550/ARXIV.2102.02079.

[2] H.-Y. Chen, C.-H. Tu, Z. Li, H.-W. Shen, and W.-L. Chao, On the importance and applicability of pretraining for federated learning, 2022. DOI: 10.48550/ARXIV.2206.11488.

[3] S. Ravuri and O. Vinyals, Classification accuracy score for conditional generative models, 2019. DOI: 10.48550/ARXIV.1905.10887.

[4] B. Xin, Y. Geng, T. Hu, et al., "Federated synthetic data generation with differential privacy," Neurocomputing, vol. 468, pp. 1–10, Jan. 2022, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.10.027