Gradient Coresets for Federated Learning

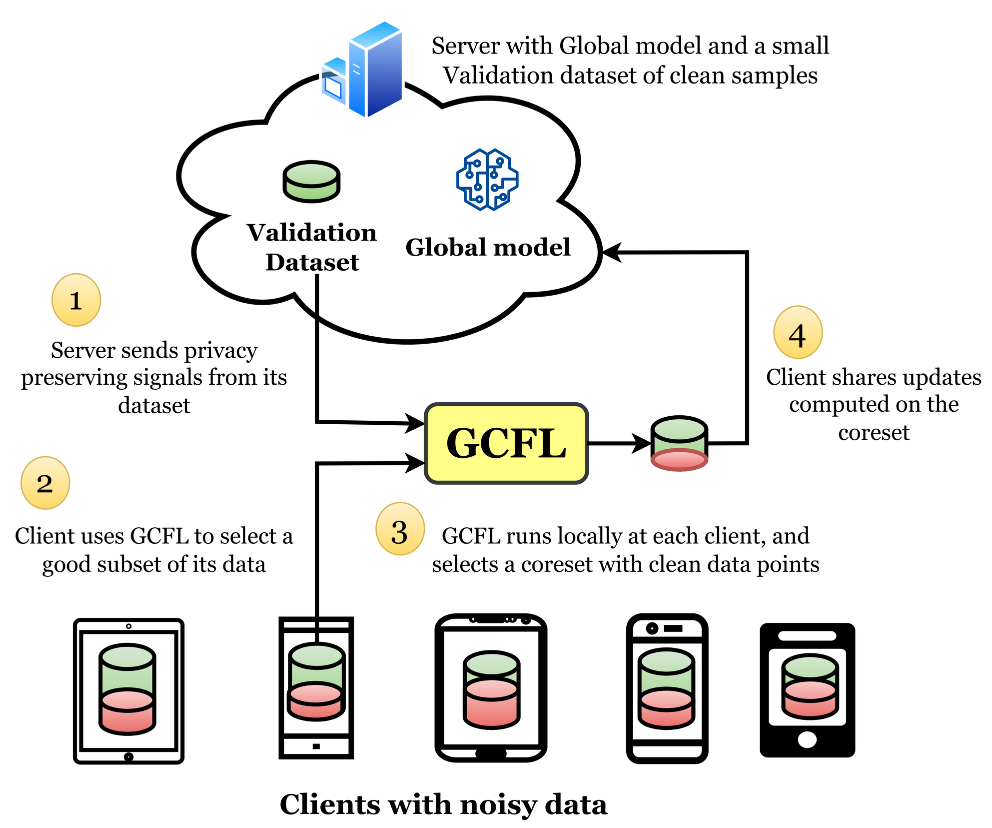
Durga Sivasubramanian, Lokesh Nagalapatti, Ganesh Ramakrishnan, Rishabh Iyer

Department of Computer Science and Engineering, IIT Bombay, UT Dallas

Introduction

- In Federated learning (FL), multiple clients collaborate to train a model with privacy-protected data.
- Clients share privacy-preserving updates iteratively during model training.
- Clients are typically edge devices, often computation and energy-deprived. Efficient solutions are crucial.
- We propose GCFL where clients compute on a data subset (coreset) and achieve energy/compute efficiency.
- Our experiments show that GCFL successfully identifies an informative subset; our models match full performance and outperform with noisy data.

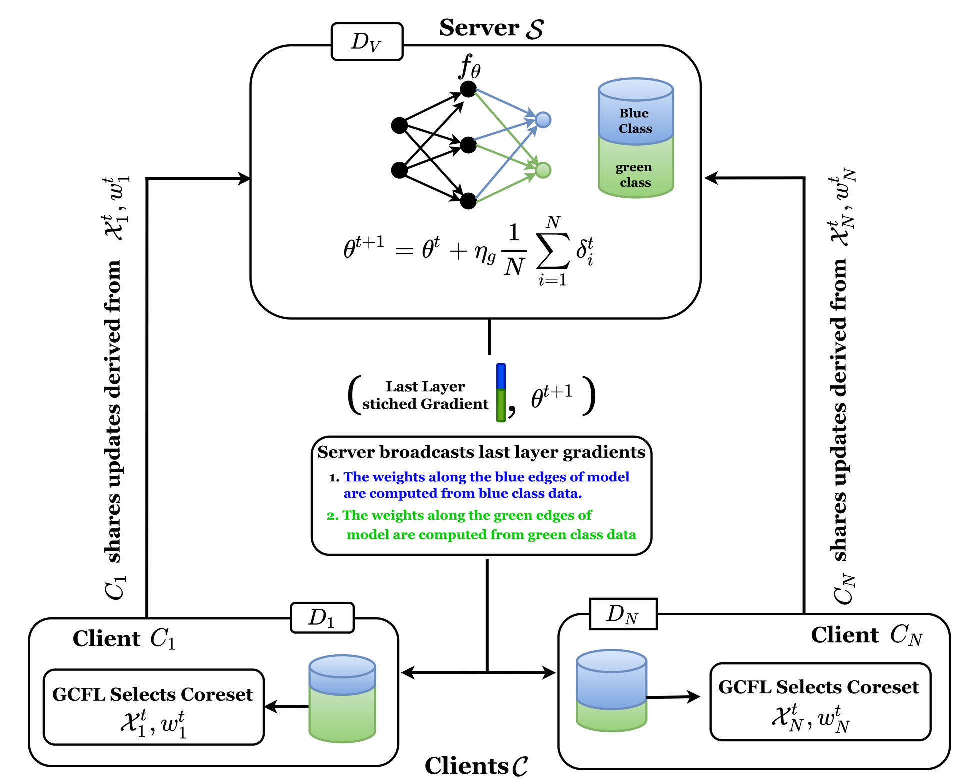
GCFL Architecture



Desiderata for coreset selection in FL

- The selected coreset at each client must match the test distribution, not just the clients'.
- The selected coreset must evolve as the model evolves; one shot selection offers poor performance.
- Assumptions in coreset approach uphold FL privacy constraints, ensuring client data confidentiality.

GCFL Solution Approach



Coreset Selection Objective

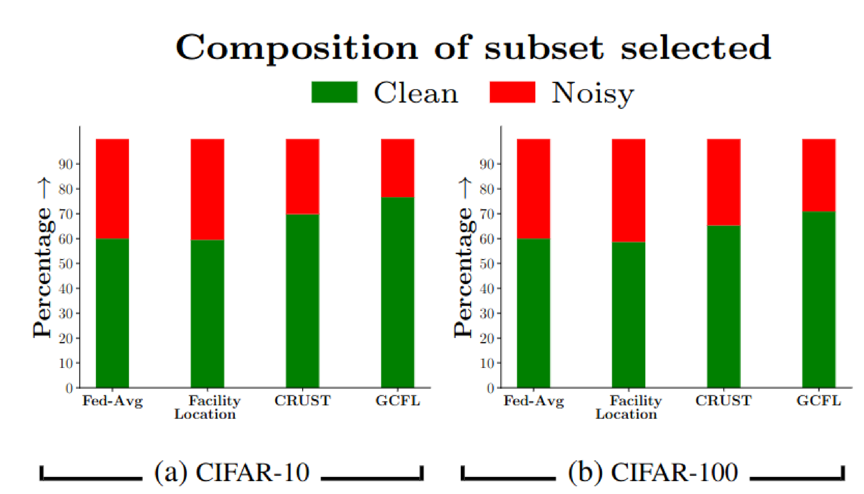
$$\underset{\mathcal{X}_i^t \subseteq D_i \text{ s.t. } |\mathcal{X}_i^t| \le b}{\operatorname{argmin}} \; \min_{\mathbf{w}_i^t} \mathrm{E}_\lambda(\mathbf{w}_i^t, \mathcal{X}_i^t) \text{ where,} \tag{1}$$

$$\mathrm{E}_\lambda(\mathbf{w}_i^t, \mathcal{X}_i^t) = \lambda \|\mathbf{w}_i^t\|^2 + \Big\| \sum_{j \in \mathcal{X}_i^t} w_{ij}^t \nabla_\theta \ell_i^j(\theta^t) - \nabla_\theta \ell_S(\theta^t) \Big\|$$

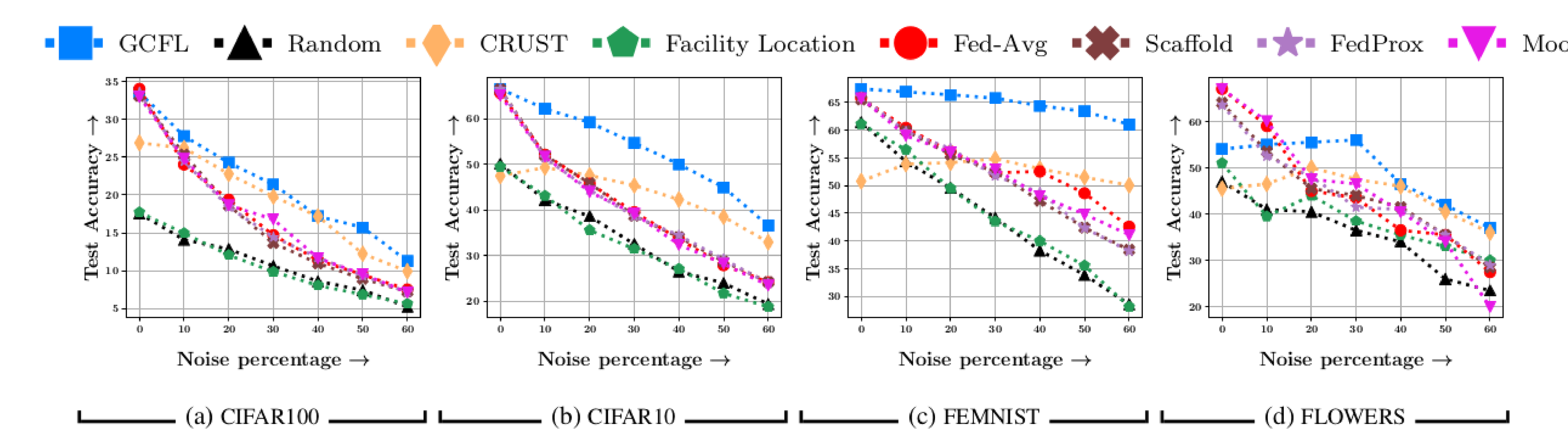
Experiments

- We consider 4 Image classification datasets: CIFAR-10, CIFAR-100, FEMNIST, Flowers.
- We compare GCFL performance with baselines under: (a) standard setting, (b) closed-set noise, (c) open-set noise, and (d) attribute noise.

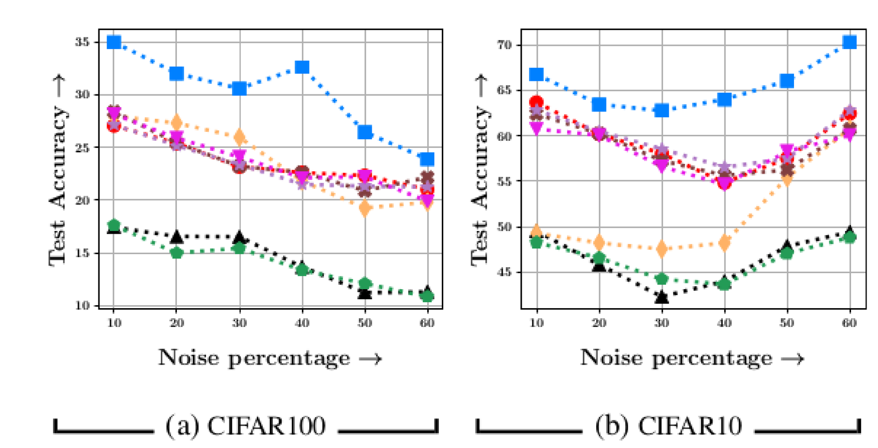
GCFL selects clean points



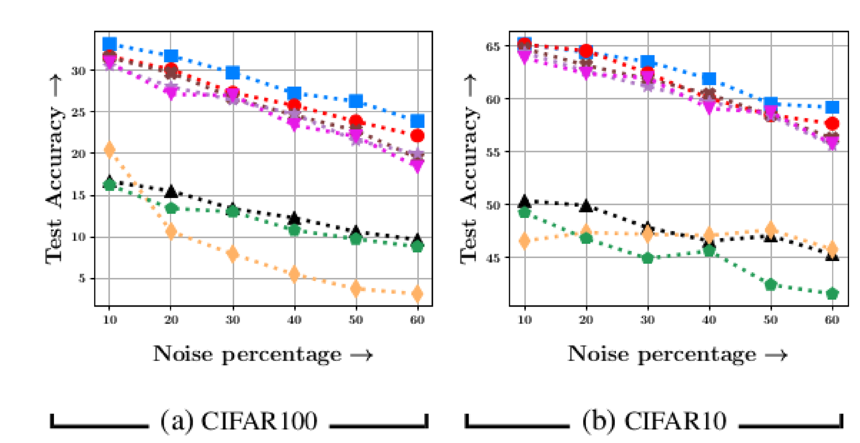
GCFL outperforms under closed set noise



GCFL outperforms under open set noise



GCFL outperforms under Attribute noise



Conclusion

- We introduced GCFL for federated learning in non-IID, and noisy data settings.
- Experimental results show GCFL outperforms state-of-the-art methods, striking the best accuracy efficiency trade-off in noise-free datasets.
- In the presence of noise, GCFL achieves significant gains compared to other FL and coreset selection baselines.