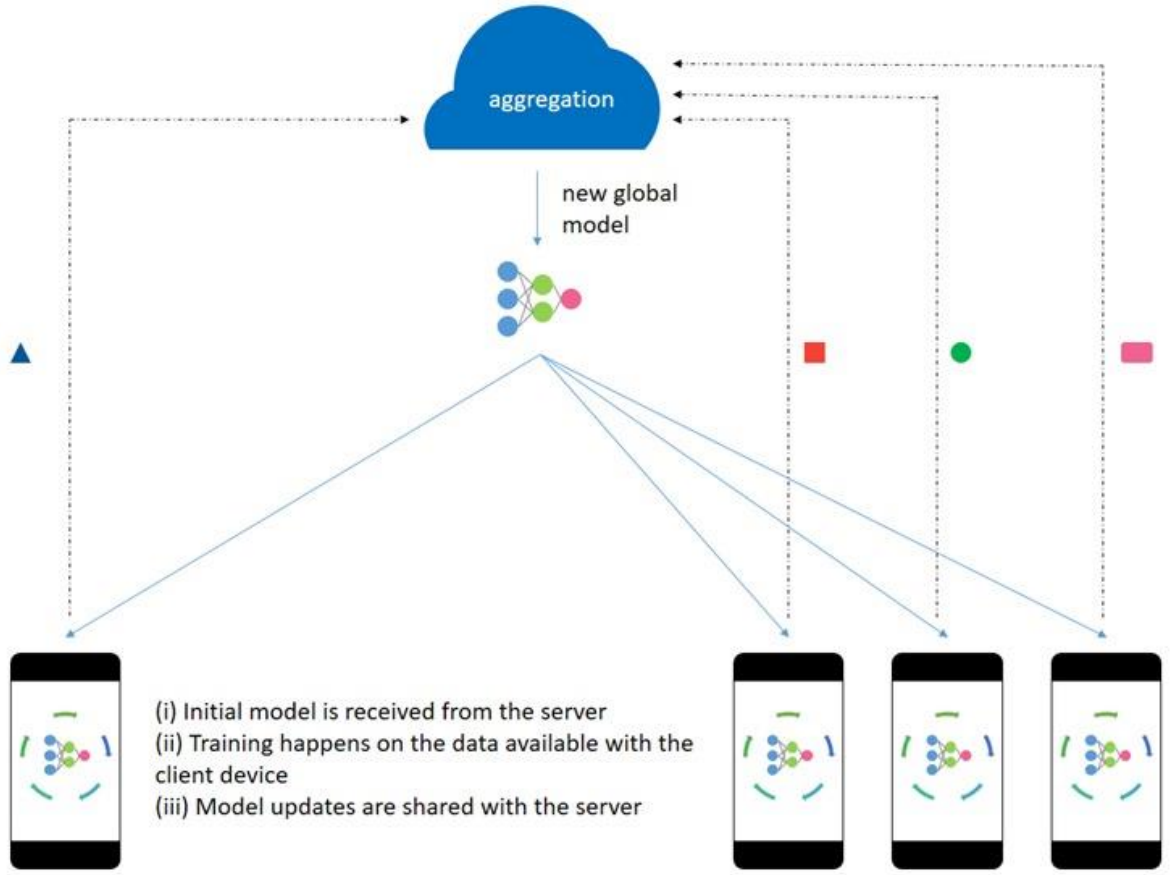
Gradient Coreset for Federated Learning

Durga Sivasubramanian*, Lokesh Nagalapatti*, Rishabh Iyer*, Ganesh Ramakrishnan*

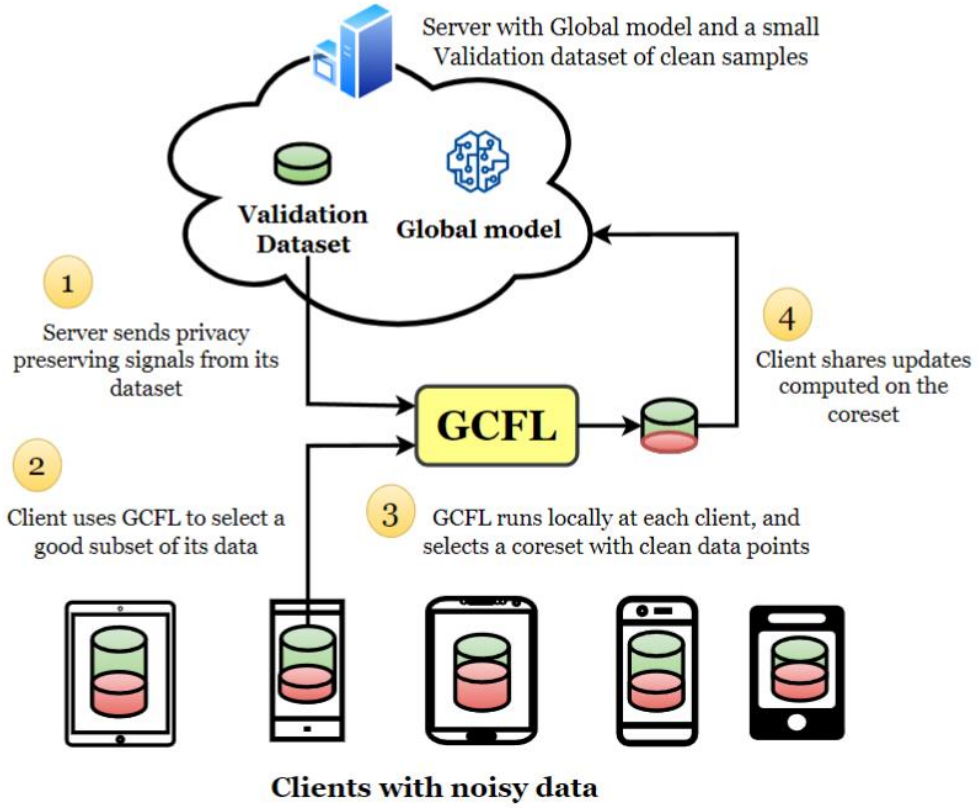
*

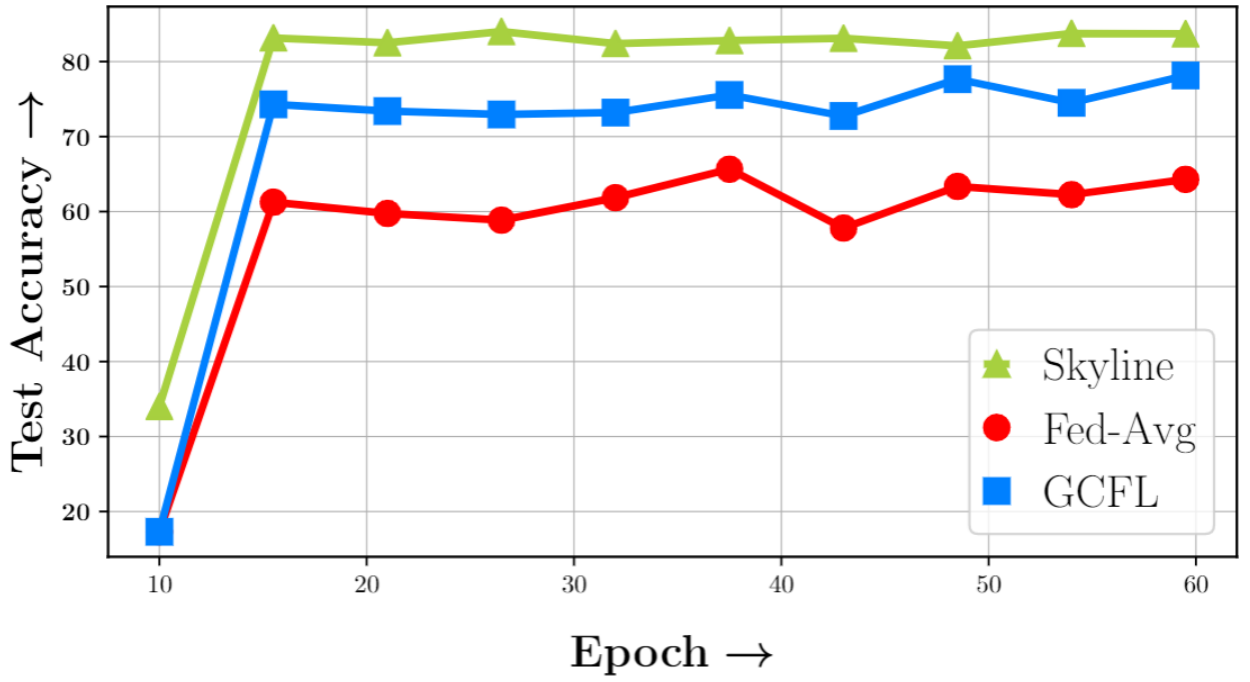
*

Federated Learning



Effect of noise in federated learning





Problem setup

We want to train machine learning model, $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ ground truth target distribution Pr_S.

However, server has $D_S \overset{\text{iid}}{\sim} Pr_S$, however $|D_S|$ is too little to train any modern machine learning models.

Therefore, the server seeks help from ***N*** clients who have enough training data,

$$D_T = \bigcup_{i=1}^{N} D_i$$

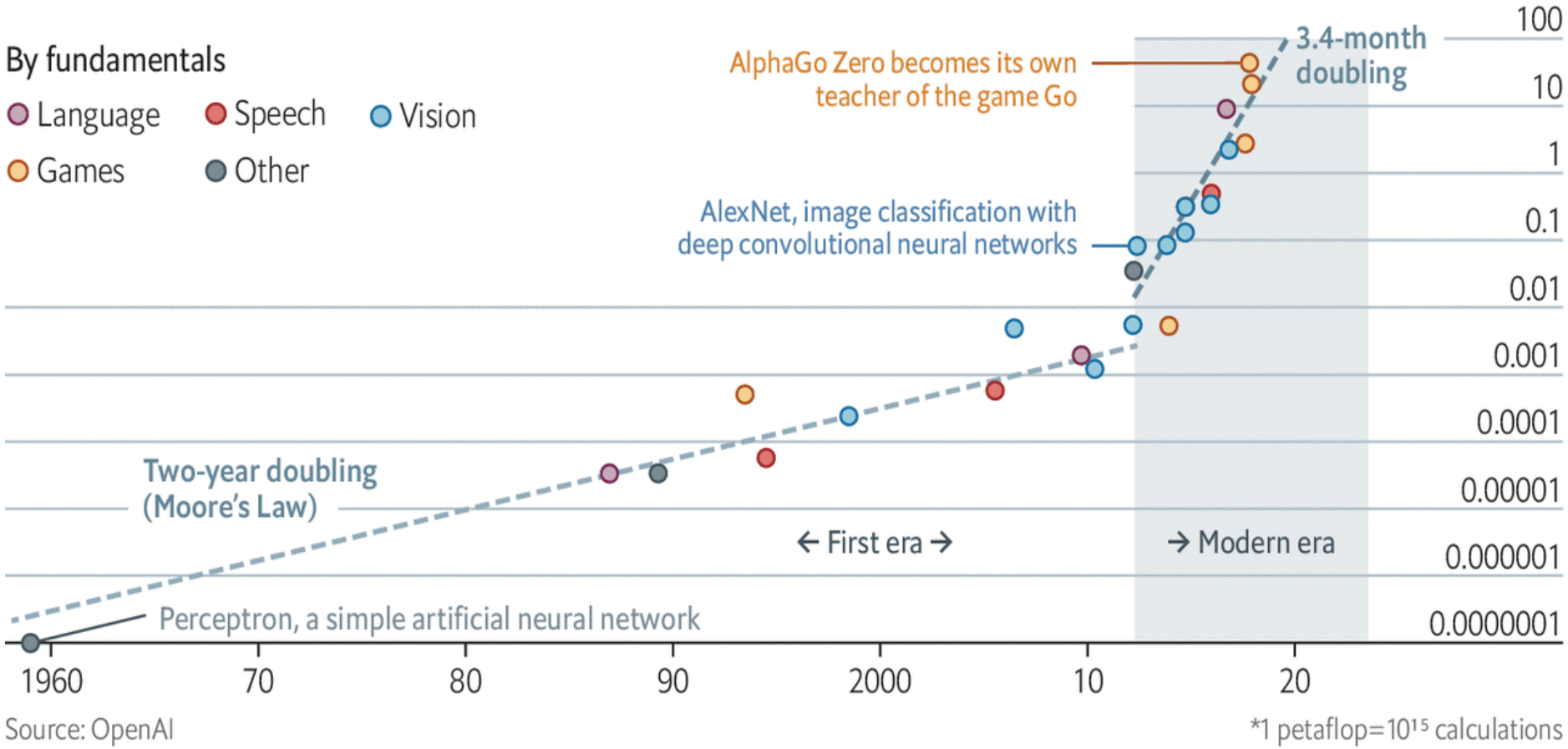
However,

$$D_T \overset{\text{iid}}{\nsim} Pr_S$$

Deep and steep

Computing power used in training AI systems

Days spent calculating at one petaflop per second*, log scale

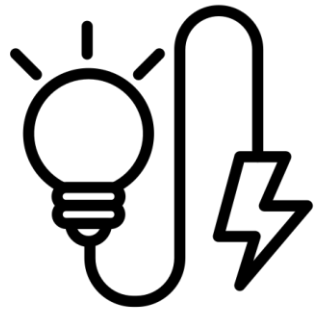


Source: OpenAI

*1 petaflop=10^{15} calculations

The Economist

Training Costs



Energy Consumption

Environmental Impact

https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem

Data explosion

In addition to noise introduced by client diversity, federated learning also inherits the problem of labeled data having redundancy. Can we intelligently subset client data to train a robust model efficiently?

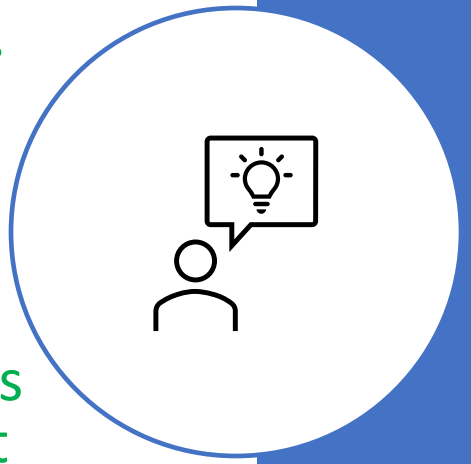
Data Subset Selection

Training on an "informative" data subset enables efficient and robust learning

To select a subset of points one need to rank points based on their suitability. Ranking could be done using a static or dynamic metric.

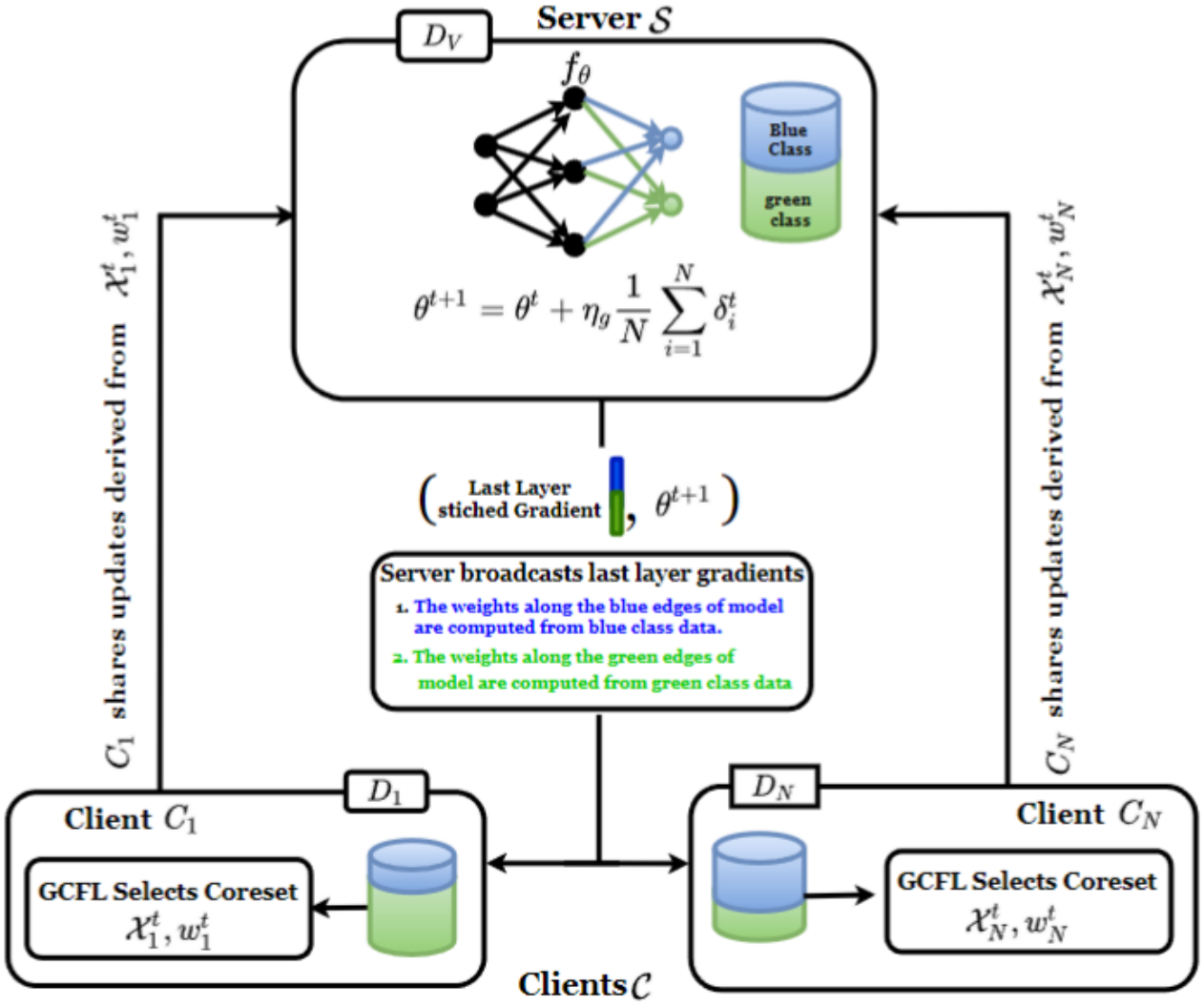
Static metric could be diversity or representation among input features

Most dynamic metrics uses loss gradient. CRUST [1] selects subsets with most representative loss gradients. GLISTER [2] selects susbet that improves validation performance the most. GRADMATCH [3] selects subset that is best able to approximate mean gradient.



1. Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models.
2. Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. 2021b. Glister: Generalization based data subset selection for efficient and robust learning. In AAAI.
3. Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021a. Grad-match: Gradient matching based data subset selection for efficient deep model training. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 5464–5474. PMLR.

Privacy preserving subset selection - GCFL



Coreset selection for Federated learning

Let loss at server's end be,

$$\ell_S = \frac{1}{|D_S|} \sum_{(x,y) \in D_S} \ell(f_\theta(x), y)$$

Then we wish to solve following optimization problem,

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

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

where, l_i^j is loss associated with j^{th} instance of i^{th} client.

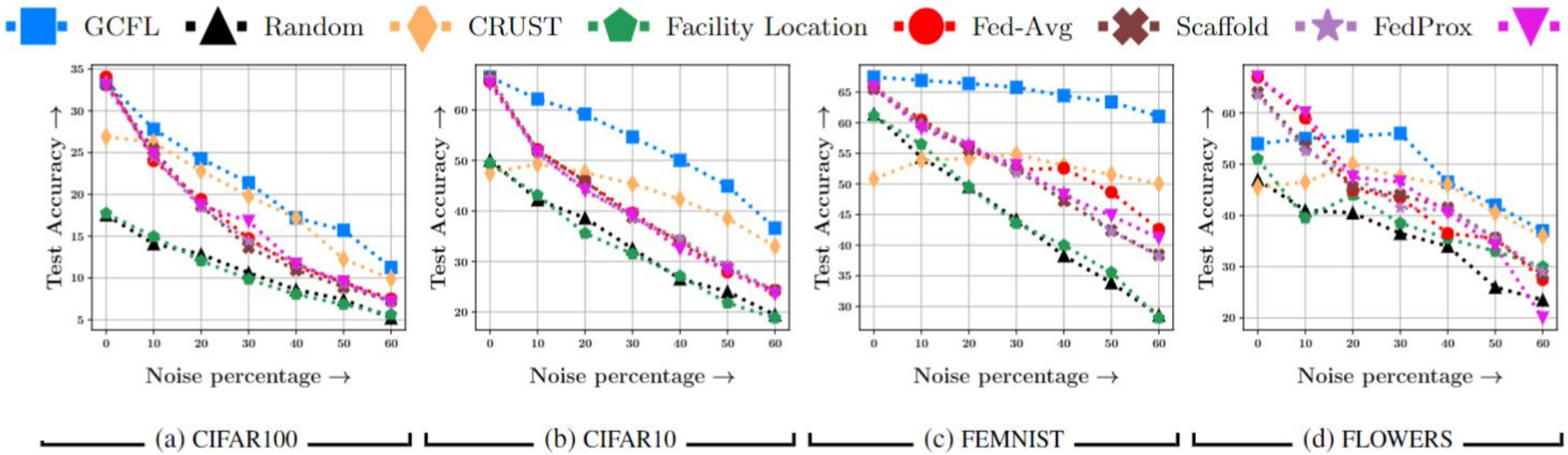
Subset Selection

The optimization problem is weakly submodular

Hence could be solved using greedy algorithm with approximation guarantees – we use orthogonal matching pursuit (OMP) algorithm

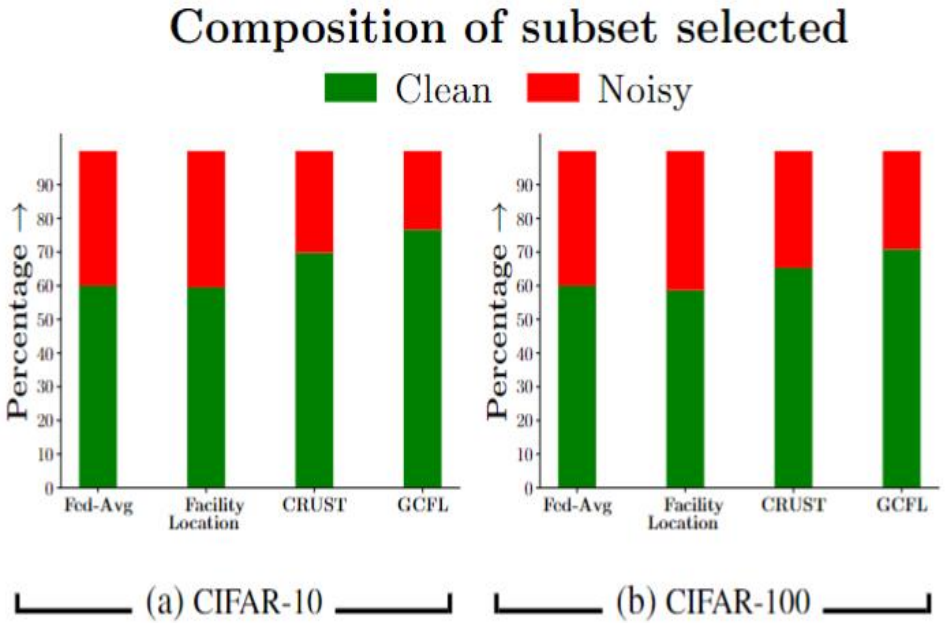
1. Find projection $r = \nabla_\theta l_i^j(\theta^t).\nabla_\theta l_s(\theta^t)$ for each $j \in D_i$ and chose the j with whose projection is maximum and add it to $\mathcal{X}_i^t$.
2. Solve linear regression problem to find w_{ij}^t for $j \in \mathcal{X}_i^t$.
3. Set $r = \nabla_\theta l_s(\theta^t) - \sum_{j \in \mathcal{X}_i^t} w_{ij}^t . \nabla_\theta l_i^j(\theta^t)$
4. Repeat the steps with new r until the $|r| < \epsilon$ or $|\mathcal{X}_i^t| < b$ (budget).
5. Return $\mathcal{X}_i^t$.

Results when clients' data is noisy

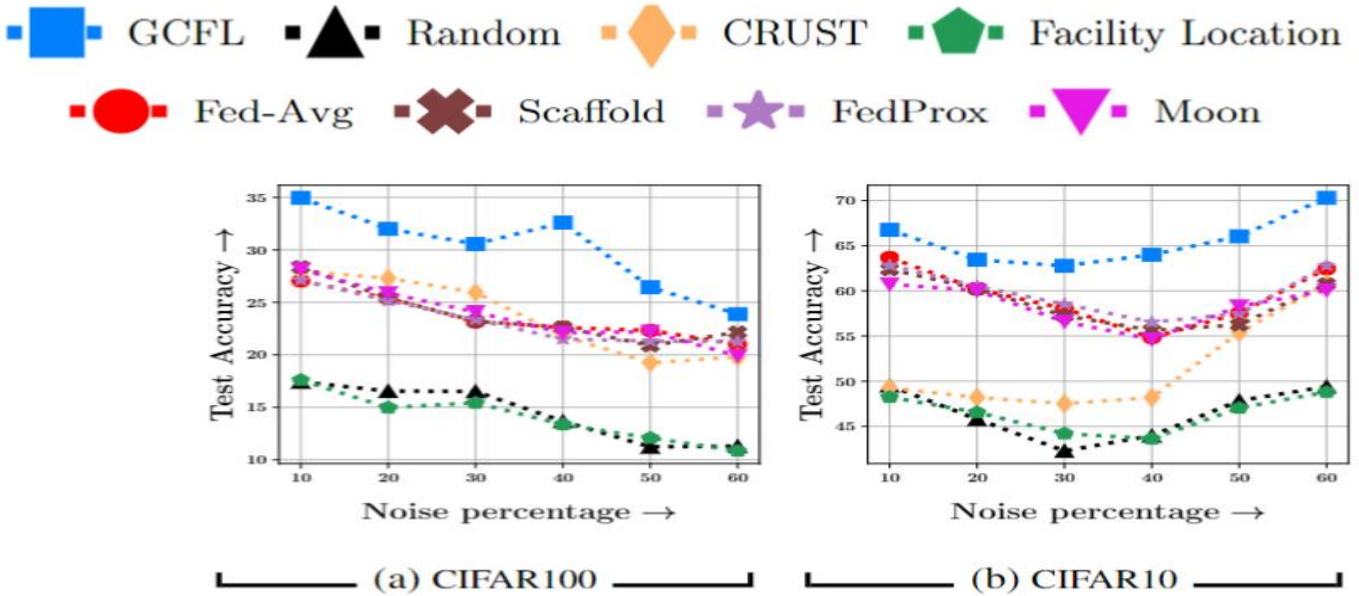


Performance comparison of GCFL and baselines with varying closed-set noise percentages. The X-axis indicates the introduced noise level, and the Y-axis shows test set accuracy. Notably, at x=0, no noise is present. Overall, GCFL outperforms the baselines, except for the flowers dataset, where subset selection hurts.

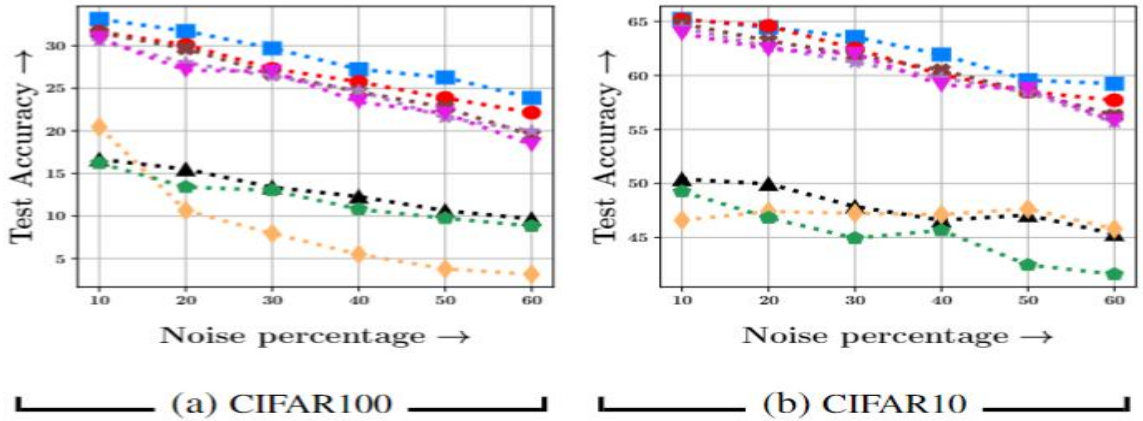
Results when clients' data is noisy



Here, we examine the number of clean points chosen for the coreset by different subset selection algorithms when trained with 40% closed-set noise. Notably, GCFL stands out by including a substantial amount of clean points in the coreset.

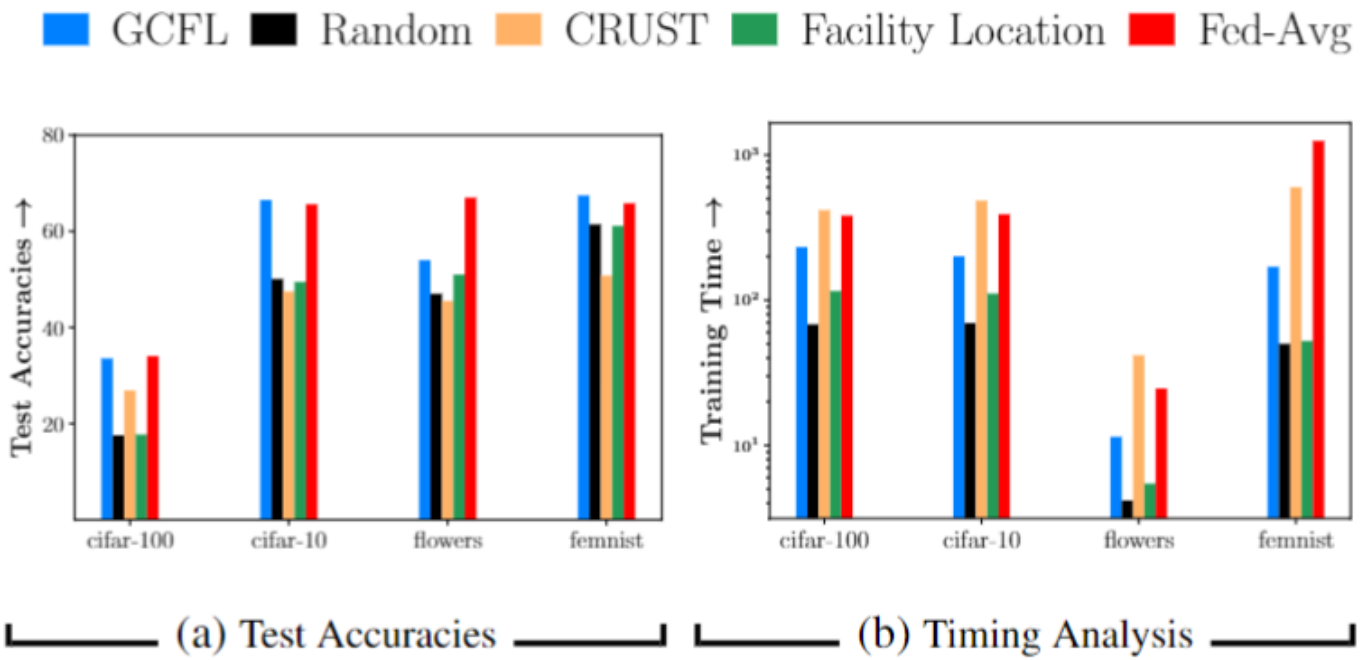


Performance of GCFL in the presence of open set noise with 10% data subset



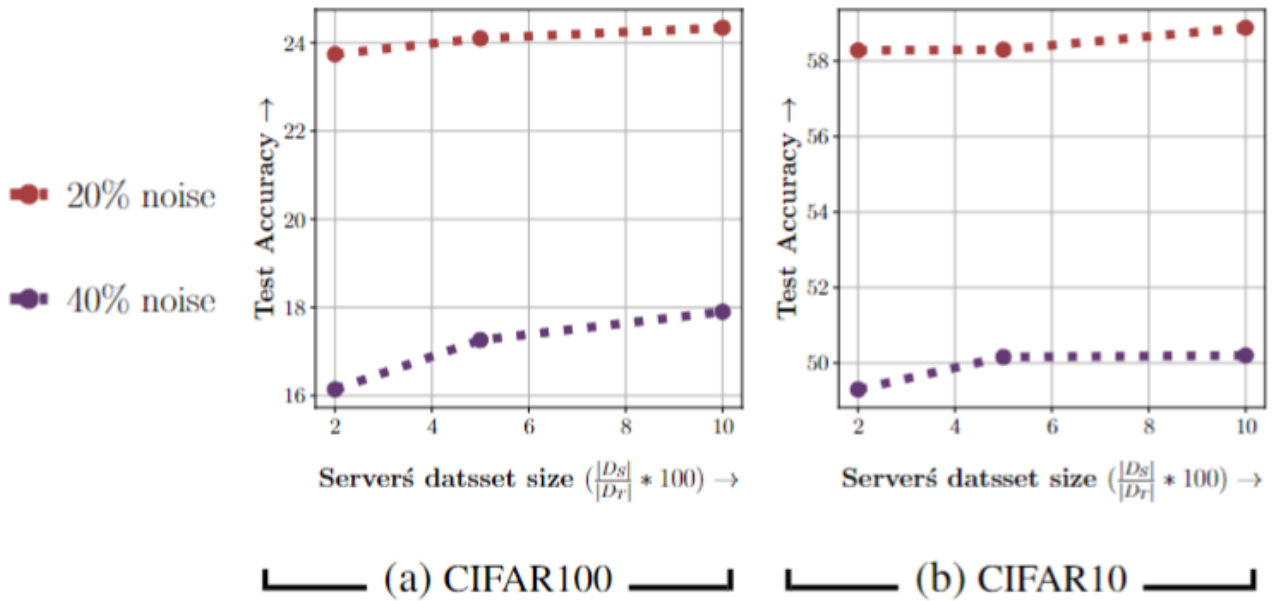
Performance of GCFL in the presence of attibute noise with 10% data subset

GCFL for improving training efficiency



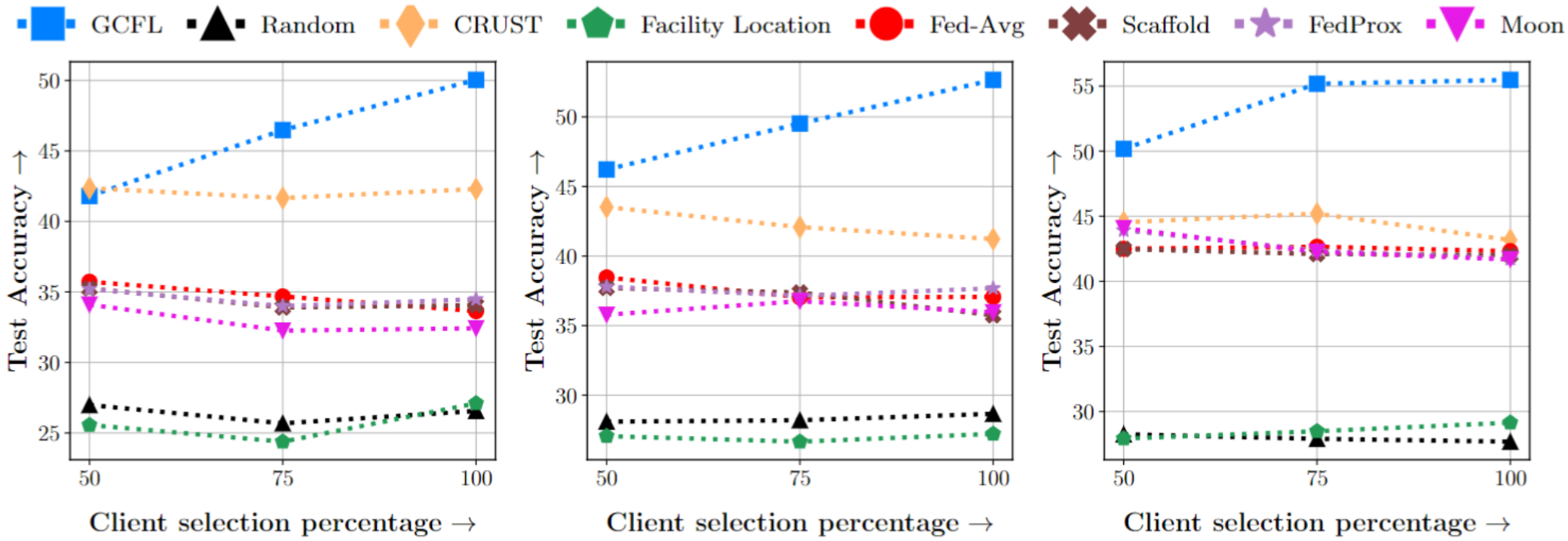
Trade-off between the training time and test accuracy on the raw datasets without any noise. We set a budget of $b = 10\%$.

Ablation on the size of $|D_S|$



Impact of server's dataset size on GCFL performance under 20%, 40% close-set noise.

Ablation on client participation



(a) 10 clients (b) 15 clients (c) 25 clients

In this experiment we vary the number of participating clients m in each round. We experiment with CIFAR-10 dataset that is injected with 40% closed-set noise. Overall we observe that GCFL performs the best.

Conclusion

We developed a gradient matching optimization algorithm for data efficient and robust training for federated learning settings.

We achieve best trade-offs between accuracy and efficiency while effectively mitigating the adverse impact of noise.

For more details, do visit our poster.