FedCS: Coreset selection for Federated Learning

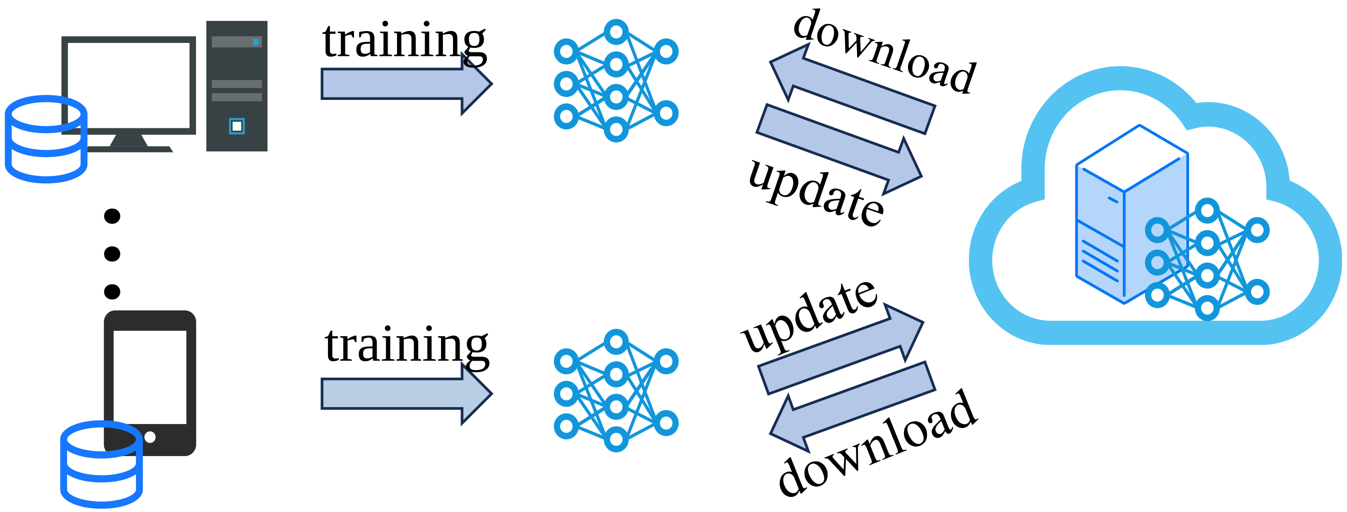
Chenhe Hao [1], Weiying Xie [*,1], Daixun Li [1],Haonan Qin [1], Hangyu Ye [1],Leyuan Fang [2], Yunsong Li [1]

[1] Xidian University, China, [2] Hunan University, China

Codes: https://github.com/xrosssaber12306/Dataset-Pruing-FL-FedCS

Federated Learning

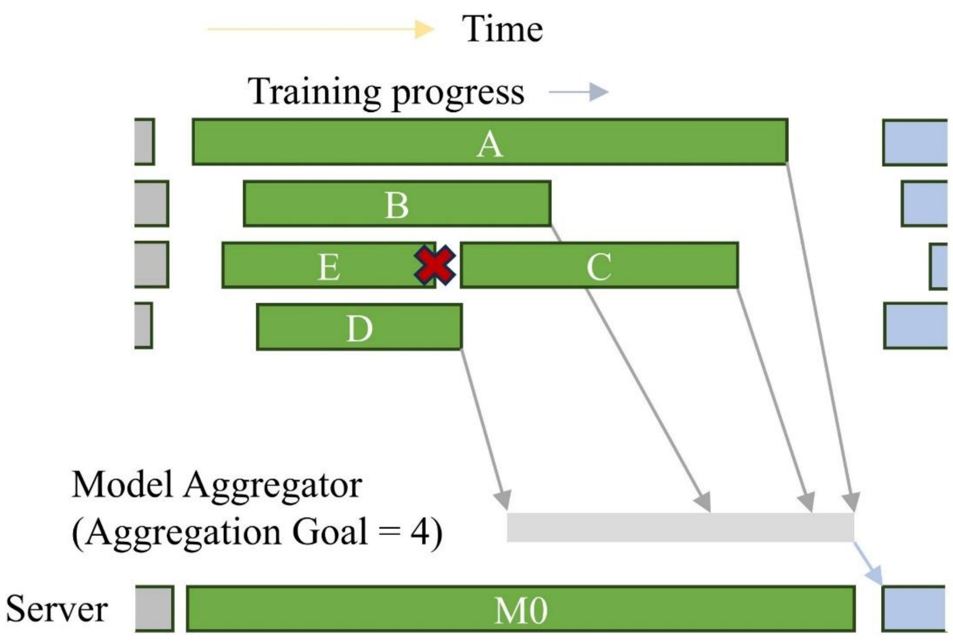
- FL is an emerging direction in **distributed machine learning** that enables jointly training a model **without sharing** the data.



Federated Learning

- **Low Training Efficiency in Federated Learning**

As datasets on clients **explosively grow**, learning from these large-scale datasets becomes progressively **time-consuming**, leading to the **training efficiency decrease** of FL.



Synchronous

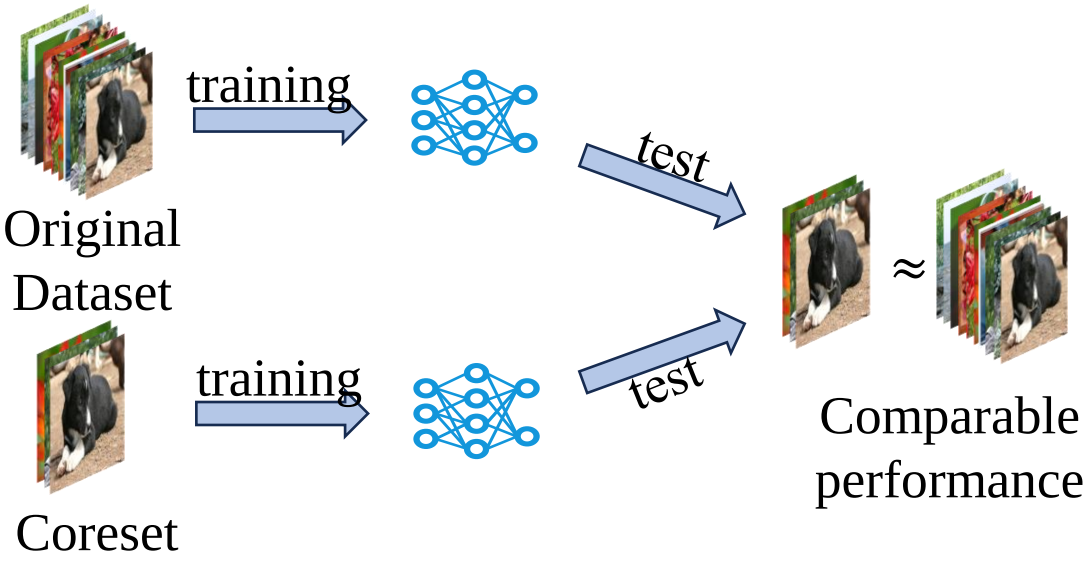
Coreset Selection

Due to the **consistency of data distribution**, a significant portion of data is **redundant.**

Objective:

$$\mathbb{E}_{\substack{(\mathbf{x},y)\sim D\\ \theta_0\sim P_{\theta_0}}}\left[\ell\left(f_{(\mathbb{U},\theta_0)}(\mathbf{x}),y\right)\right] \simeq \mathbb{E}_{\substack{(\mathbf{x},y)\sim D\\ \theta_0\sim P_{\theta_0}}}\left[\ell\left(f_{(\mathbb{S},\theta_0)}(\mathbf{x}),y\right)\right]$$

- Select a subset of images in a full dataset without performance drop

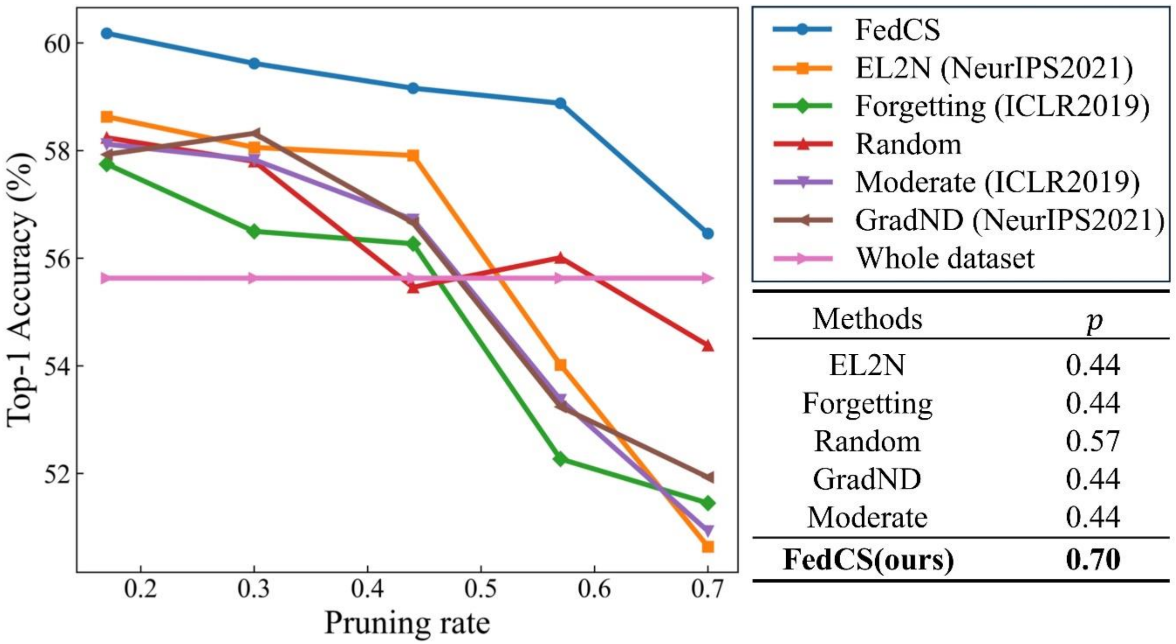


Coreset Selection: An efficient tool to improve training efficiency.

Motivation:

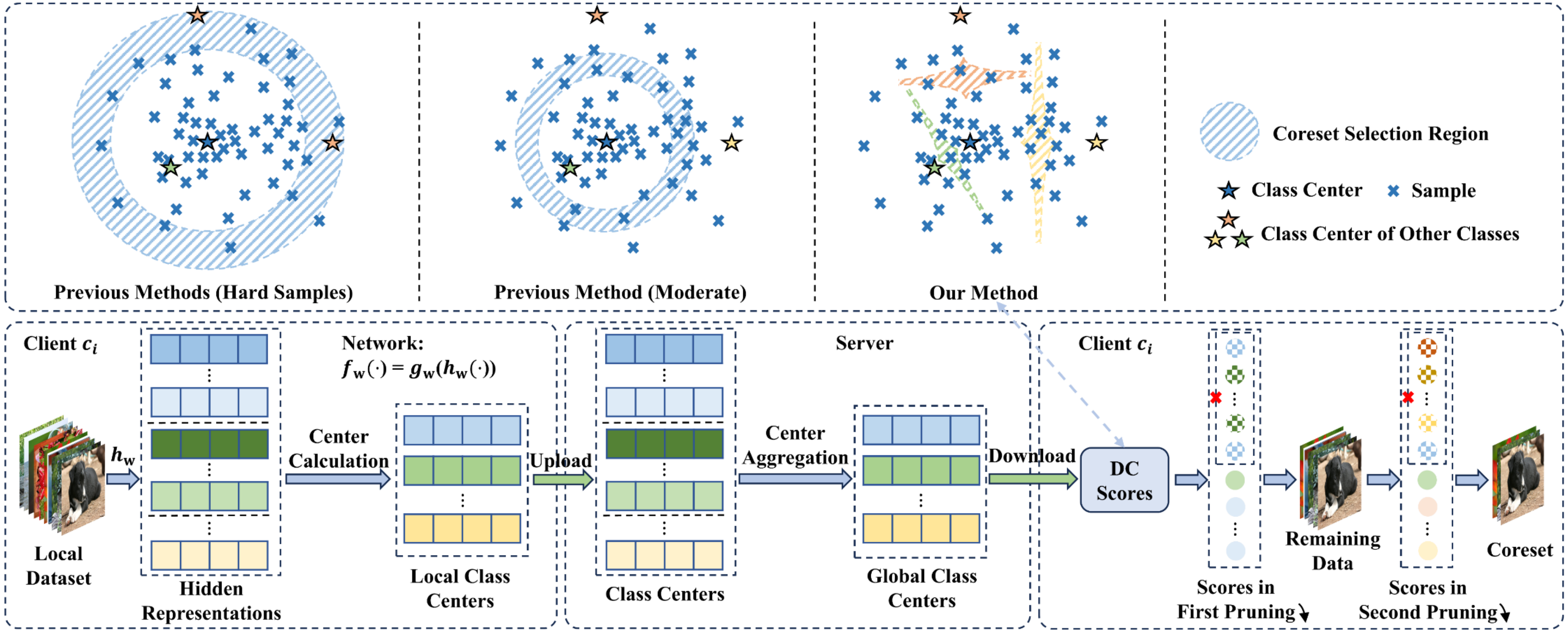
- **Existing coreset selection methods applied to FL ignore heterogeneity of data distribution.**

The Coreset Selection work in FL **lacks** an effective lossless pruning criterion. They usually pay **less attention** to the **heterogeneity of the data distribution.** As a result, applying such a existing coreset selection criterion to FL will lead to exacerbation of data distribution imbalance.



Methods	p
EL2N	0.44
Forgetting	0.44
Random	0.57
GradND	0.44
Moderate	0.44
FedCS(ours)	**0.70**

Methodology:



overview of our proposed FedCS

Methodology:

Class Center Aggregation

$$\{z_i^k = \frac{\sum_{j=1}^{n_i} \mathbb{I}(y_{ij} = k) z_{ij}}{\sum_{j=1}^{n_i} \mathbb{I}(y_{ij} = k)}\}_{k \in [\mathcal{K}]} \Longrightarrow \{z^k = \mathrm{median}(z_i^k)_{i=1}^N\}_{k \in [\mathcal{K}]}$$

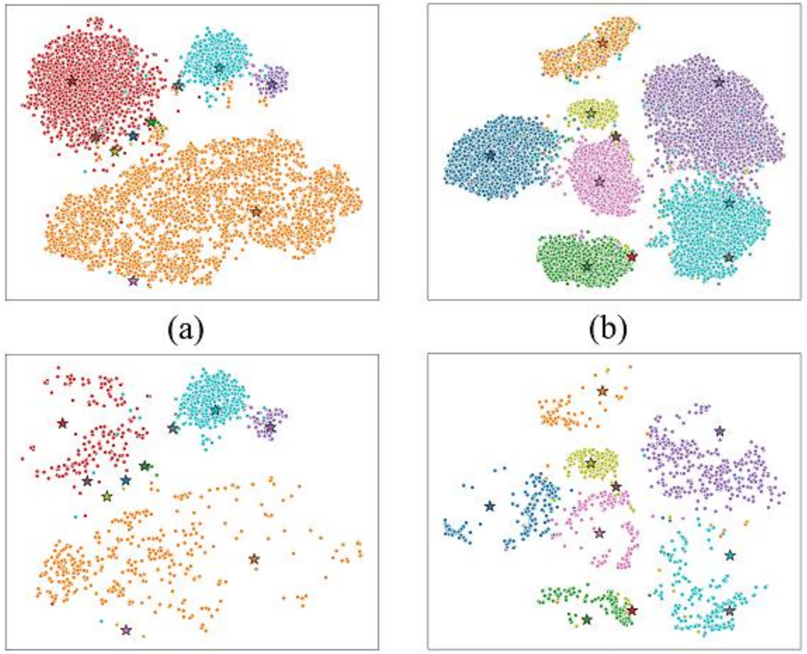
Local class center — Global class center

Class center aggregation avoids important samples from overlooking caused by discrepancy between client sample distributions and global sample distribution.

Methodology:

DC Score

Choose the samples around **decision boundary.**


(a) (b) (c) (d)

$$s_{ij} = \left| \min_{k \neq y_{ij}} \{d_{ij}^{k}\}_{k \in [\mathcal{K}]} - d_{ij}^{y_{ij}} \right|$$

$$d_{ij}^{k} = \left\| z_{ij} - z^{k} \right\|_2$$

Methodology:

Double Pruning

First perform **high-ratio pruning** on “Large-capacity Class” alone with **higher** pruning ratio p_i^f .

$$\mathrm{Top}-M_f\left\{D_i^{pf}, S_i^{pf}\right\} \leftarrow \{(\mathbf{x}_{ij}^L, y_{ij}^L), s_{ij}\}_{j=1}^{M_f}$$

$$D_i^r = D_i \backslash D_i^{pf}$$

Finally perform **low-ratio pruning** on remain samples alone with **lower** pruning ratio p_i^l.

$$\mathrm{Top}-M_f\left\{D_i^{pf}, S_i^{pf}\right\} \leftarrow \{(\mathbf{x}_{ij}^L, y_{ij}^L), s_{ij}\}_{j=1}^{M_f}$$

$$D_i^r = D_i \backslash D_i^{pf}$$

Convergence Analysis:

$$\mathbb{E}\left[F\left(\mathbf{w}^{(T)}, D^*\right)\right] - F^* \leq \frac{1}{(T+\gamma)}\left[\frac{4L(32\tau^2G^2 + \sigma^2/m}{3\mu^2\bar{\rho}} + \frac{8L^2\Gamma}{\mu^2} + \frac{L\gamma\left\|\overline{\mathbf{W}}^{(0)} - \mathbf{w}^*\right\|^2}{2}\right] + \frac{8L\Gamma}{3\mu}(\frac{\tilde{\rho}}{\bar{\rho}} - 1)$$

Compared to other methods which select hard samples, FedCS **accelerates convergence** and gets final solution with **lower error** by **reducing the Degree of Non-iid Γ.**

FedCS: Coreset Selection for Federated Learning

Experiments:

Cifar10 on ResNet-18

Methods	$\alpha = 0.1, p_i^l = 0.1$					$\alpha = 1.0, p_i^l = 0.1$				
	$p_i^f = 0.1$	$p_i^f = 0.3$	$p_i^f = 0.5$	$p_i^f = 0.7$	$p_i^f = 0.9$	$p_i^f = 0.1$	$p_i^f = 0.3$	$p_i^f = 0.5$	$p_i^f = 0.7$	$p_i^f = 0.9$
EL2N	58.63±0.79	58.06±0.93	57.91±1.27	54.02±1.33	50.64±1.43	66.55±0.61	64.28±0.71	60.57±0.73	55.72±0.80	53.07±0.94
Moderate	58.12±0.63	57.83±0.68	56.72±0.72	53.37±0.75	50.93±0.82	66.18±0.37	64.78±0.48	63.40±0.54	60.35±0.60	56.65±0.67
GM	58.48±0.51	57.97±0.58	56.82±0.61	57.53±0.66	55.52±0.71	66.40±0.25	65.17±0.31	63.59±0.34	60.79±0.44	57.06±0.54
GradND	57.93±0.70	58.32±1.03	56.66±1.14	53.24±1.22	51.93±1.27	66.29±0.54	64.21±0.55	60.83±0.77	56.20±0.87	52.06±0.91
Forgetting	57.75±0.74	56.50±0.97	56.27±1.25	52.27±1.33	51.45±1.46	66.18±0.64	62.56±0.88	62.28±0.98	59.15±1.03	54.76±1.11
Random	58.24±0.86	57.80±1.07	55.46±1.44	56.01±1.64	54.38±2.04	66.52±0.67	65.45±0.78	62.17±0.79	61.02±1.15	58.18±1.48
FedCS(ours)	**60.18**±0.73	**59.62**±0.79	**59.16**±0.81	**58.88**±0.86	**56.46**±0.88	**66.98**±0.44	**66.11**±0.49	**63.82**±0.60	**61.64**±0.71	**58.42**±0.76
Whole Dataset	55.63±0.68	55.63±0.68	55.63±0.68	55.63±0.68	55.63±0.68	66.04±0.24	66.04±0.24	66.04±0.24	66.04±0.24	66.04±0.24

Better performance & higher compression ratio

Experiments:

Ablation Study

Cifar10 on ResNet-18

Methods	CA	DP	$\alpha = 0.1, p_i^l = 0.1$				$\alpha = 1.0, p_i^l = 0.1$			
			$p_i^f = 0.3$ $p_i = 0.30$	$p_i^f = 0.5$ $p_i = 0.44$	$p_i^f = 0.7$ $p_i = 0.57$	$p_i^f = 0.9$ $p_i = 0.70$	$p_i^f = 0.3$ $p_i = 0.32$	$p_i^f = 0.5$ $p_i = 0.47$	$p_i^f = 0.7$ $p_i = 0.61$	$p_i^f = 0.9$ $p_i = 0.76$
FedCS w/o CA, DP			57.41±1.01	56.56±1.25	52.76±1.45	51.55±1.57	63.80±0.72	61.28±0.86	59.32±1.01	55.14±1.12
FedCS w/o DP	✓		57.77±1.09	57.43±1.16	55.37±1.30	53.16±1.45	64.23±0.64	61.67±0.75	60.01±0.81	55.75±0.93
FedCS w/o CA		✓	58.22±0.93	58.11±1.02	56.26±1.11	53.78±1.19	64.90±0.57	63.37±0.64	61.28±0.73	57.51±0.82
FedCS	✓	✓	**59.62±0.79**	**59.16±0.81**	**58.88±0.86**	**56.46±0.88**	**66.11±0.49**	**63.82±0.60**	**61.64±0.71**	**58.42±0.76**
Whole Dataset			55.63±0.68	55.63±0.68	55.63±0.68	55.63±0.68	66.04±0.24	66.04±0.24	66.04±0.24	66.04±0.24

Thank you