To Store or Not? Online Data Selection for Federated Learning with Limited Storage

Chen Gong[1], Zhenzhe Zheng[1], Yunfeng Shao[2], Bingshuai Li[2], Fan Wu[1], Guihai Chen[1]

Shanghai Jiao Tong University[1]

Huawei Noah’s Ark Lab[2]

NOAH'S ARK LAB

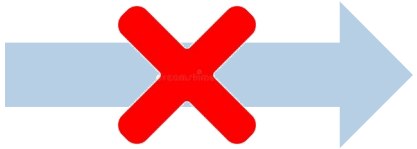
From Big Data to Deep Knowledge

Mobile Network needs ML models

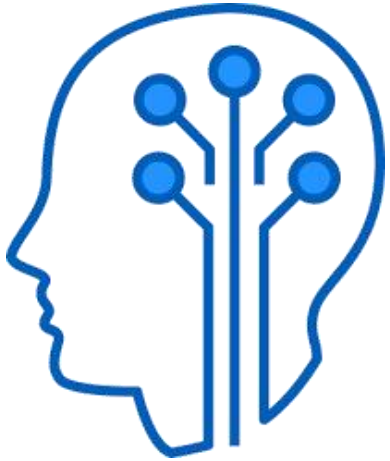
Large Scale

Complex Correlation

Low Latency

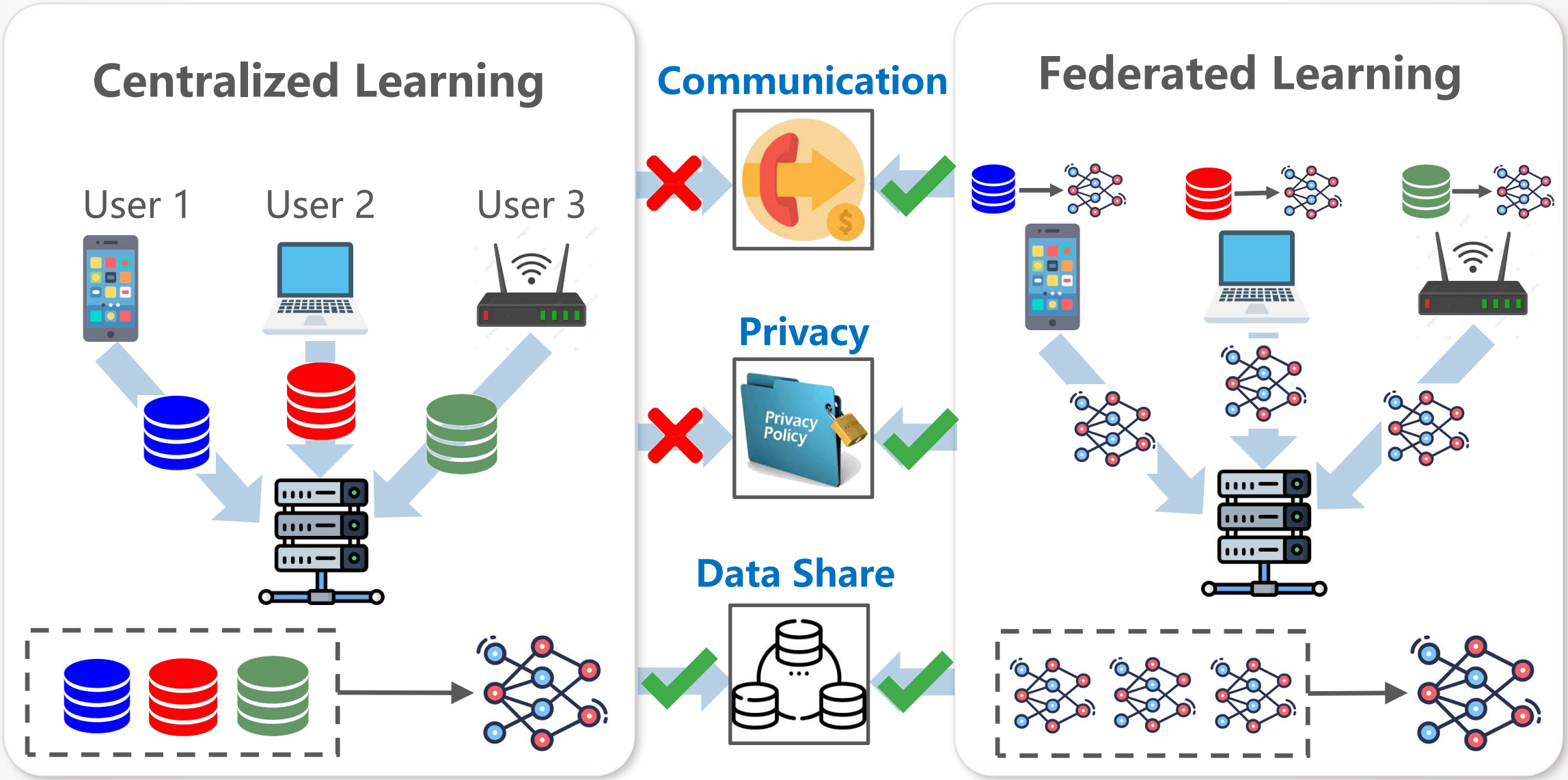


Human-in-the-loop Approach



Machine Learning Models

ML Model Training needs FL



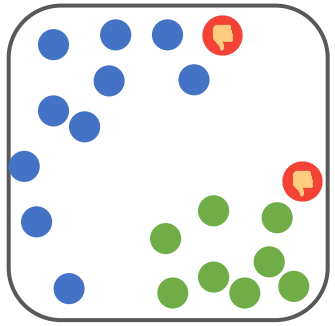
Device Properties Degrade FL Performance

Device Properties

Streaming Data

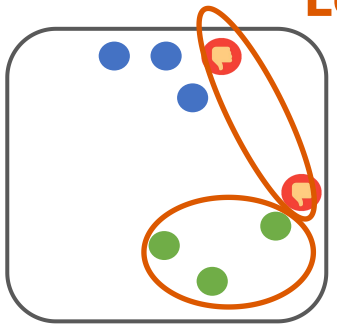
Limited Storage

Locally Stored Data



True data distribution

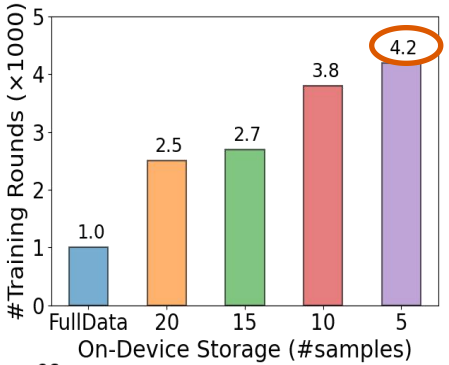
Low-quality



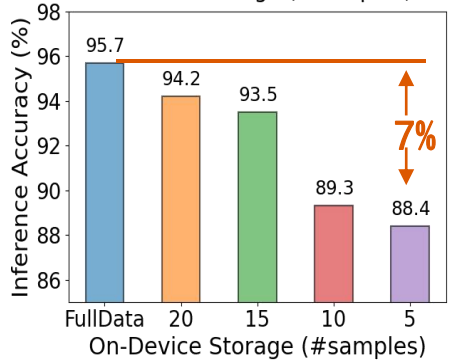
On-device stored data

Biased

Performance Degradation



Training Time increases 4.2×



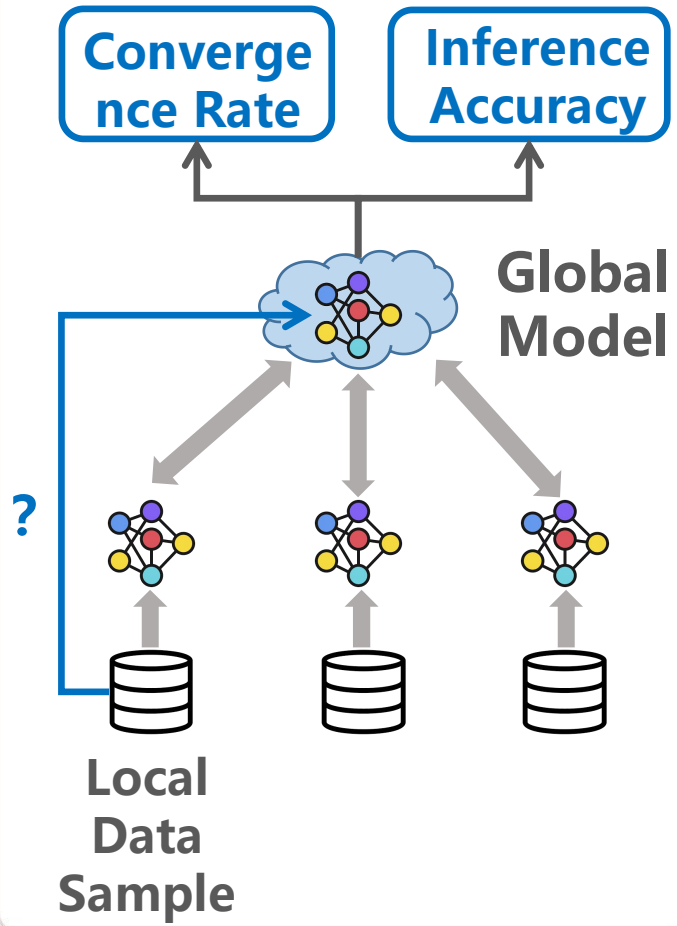
Final Acc drops 7%

Problem

Filter valuable on-device data to simultaneously improve training convergence and inference accuracy of FL model

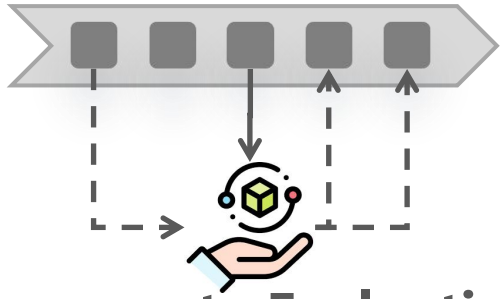
Three Key Challenges

Theoretical Guarantee



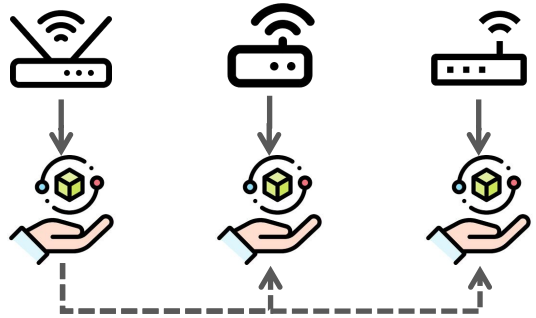
Lack of Multi-Dim Information

Temporal Information



Inaccurate Evaluation

Spatial Information

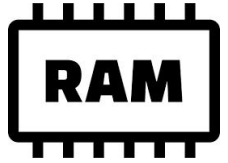


Redundant Stored Data

Constrained Resource

Computation Cost

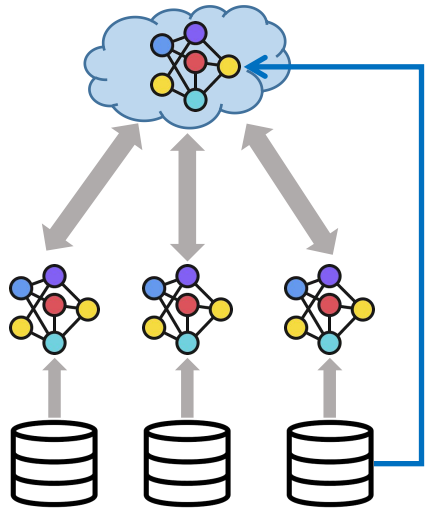
Evaluation Delay



Memory Footprint

Theoretical Analysis

Local Data Impact



Convergence Rate

$$\underbrace{F(w_{\text{fed}}^{t-1}) - F(w_{\text{fed}}^{t})}_{\text{global loss reduction}} \geqslant \sum_{c \in C_t} \sum_{i=0}^{m-1} \sum_{(x,y) \in B_c} \Big[\underbrace{-\alpha_c \parallel \nabla_w l(w_c^{t,i}, x, y) \parallel_2^2}_{\text{term 1}} + \beta_c \underbrace{\langle \nabla_w l(w_c^{t,i}, x, y), \nabla_w F(w_{\text{fed}}^{t-1}) \rangle}_{\text{term 2}} \Big],$$

Global loss reduction

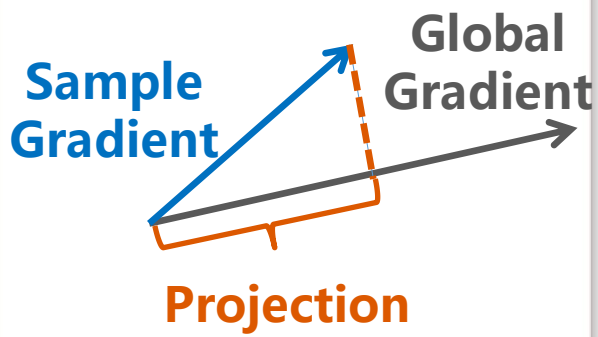
Dominant Term

Inference Accuracy

$$\underbrace{\parallel w_{\text{fed}}^{t} - w_{\text{cen}}^{mt} \parallel_2}_{\Delta\text{model of CL and FL}} \leqslant (1+\eta L)^m \parallel w_{\text{fed}}^{t-1} - w_{\text{cen}}^{m(t-1)} \parallel_2 + \sum_{c \in C_t} \zeta_c^t \left[\eta \sum_{i=0}^{m-1} (1+\eta L)^{m-1-i} G_c(w_c^{t,i}) \right],$$

Dominant Term

Common Term

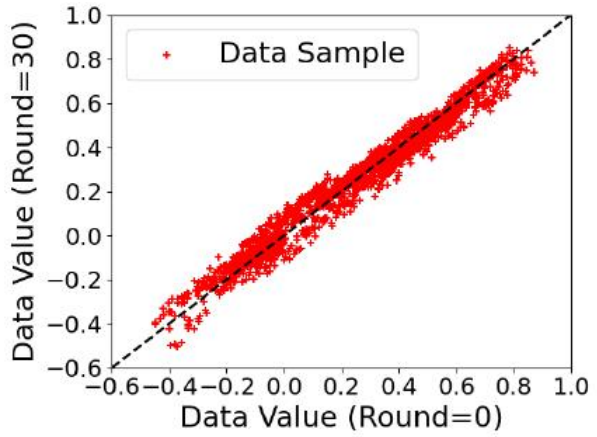


Intuition

- Preserve **personal information**
- Reduce global data **heterogeneity**

On-Device Data Selection

Lack of Latest Global Model



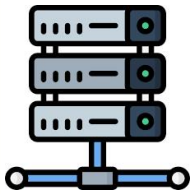
(b) Round 0 vs Round 30

Data **value remains stable** across few training rounds

Lack of Accurate Global Gradient

Local Gradient Estimator

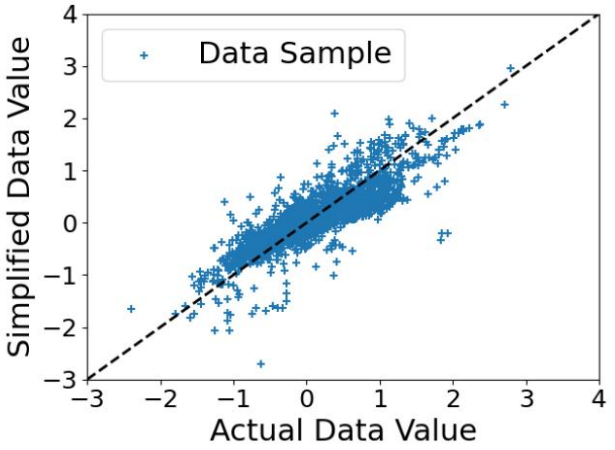
$$\hat{g}_c \leftarrow \frac{n-1}{n}\hat{g}_c + \frac{1}{n}\nabla_w l(w_{\text{fed}}^{t_{c,\text{last}}-1}, x, y).$$



Global Gradient Estimator

$$\hat{g}^t \leftarrow \hat{g}^{t-1} + \sum_{c \in C_t} \zeta_c (\hat{g}_c - \hat{g}_c^{t_{\text{last}}}),$$

Simplification Version

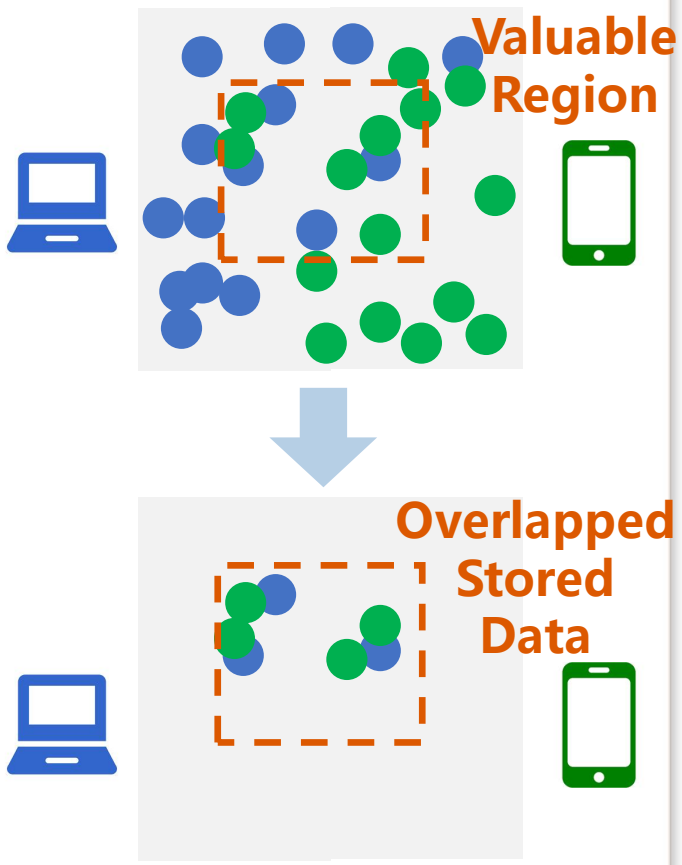


(c) Full model vs Last layer

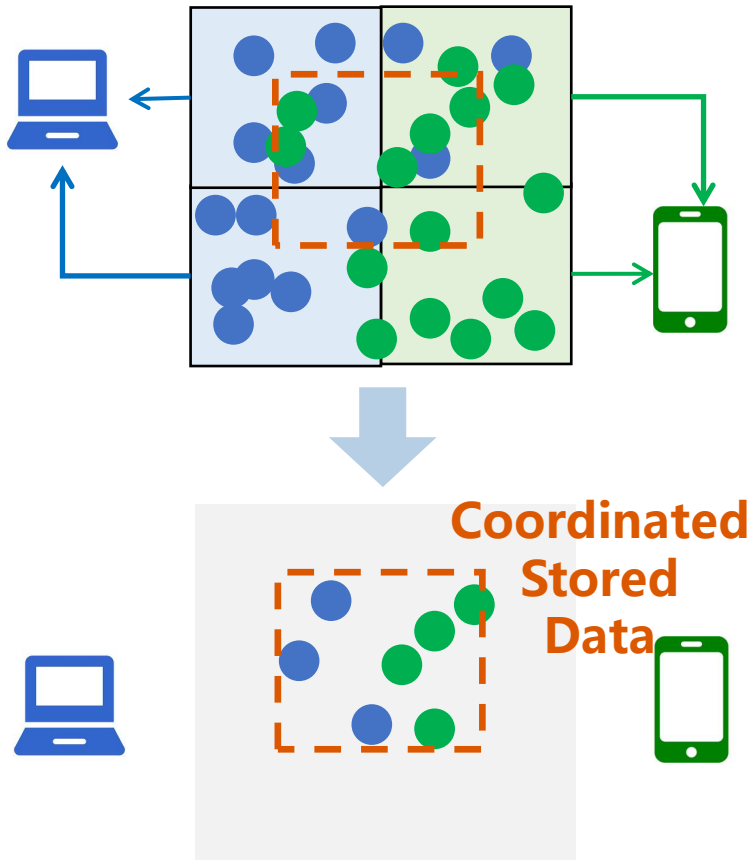
Partial model gradient is able to reflect the **overall tendency**

Cross-Device Data Storage

Overlapped Stored Data



Coordination Policy



Problem Formulation

- **Efficient Data Selection**
- Redundant Label Assignment
- Limited Storage
- Unbiased global data distribution

$$\max_{D} \quad \sum_{c\in C}\sum_{y\in Y} D_{c,y}\cdot V_{c,y} = \| DV \|_1,$$

$$\begin{aligned}
\text{s.t.} \quad & \| D_y^{\mathrm{T}} \|_0 \geqslant n_y^{\text{label}}, && \forall y \in Y,\\
& \| D_c \|_0 \leqslant n_c^{\text{client}}, && \forall c \in C,\\
& \| D_c \|_1 = |B_c|, && \forall c \in C,\\
& \frac{\sum_{c\in C} D_{c,y}}{\| B \|_1} = \frac{\sum_{c\in C} V_{c,y}}{\| V \|_1}, && \forall y \in Y.
\end{aligned}$$

Experiments

Learning Tasks and Models

Synthetic
Image Classification
Activity Recognition
Traffic Classification

Datasets	Models	#Samples	#Labels	#Devices	n_y^{label}	$\frac{\lvert C_t \rvert}{\lvert C \rvert}$	η	m	$\lvert B_c \rvert$
Synthetic Dataset [31]	LogReg	1,016,442	10	200	5	5%	$1e^{-4}$	5	10
Fashion-MNIST [32]	LeNet [58]	70,000	10	50	5	10%	$1e^{-3}$	5	5
HARBOX [33]	Customized DNN	34,115	5	120	5	10%	$1e^{-3}$	5	5
Industrial Dataset	Customized CNN	37,853	20	30	5	20%	$5e^{-3}$	5	10

Videos

Games

Downloading

Communication

Experiments

Overall Performance

Task	Model Training Speedup							
	RS	HL	GN	FB	SLD	ODE-Exact	ODE-Est	FD
ST	1.0×	–	4.87×	–	4.08×	9.52×	5.88×	2.67×
IC	1.0×	–	–	–	–	1.35×	1.20×	1.01×
HAR	1.0×	–	–	–	–	2.22×	1.55×	4.76×
TC	1.0×	–	–	–	–	2.51×	2.50×	3.92×
Task	**Inference Accuracy**							
ST	79.56%	78.44%	83.28%	78.56%	82.38%	87.12%	82.80%	88.14%
IC	71.31%	51.95%	41.45%	60.43%	69.15%	72.71%	72.70%	71.37%
HAR	67.25%	48.16%	51.02%	48.33%	56.24%	73.63%	70.39%	77.54%
TC	89.3%	69.00%	69.3%	72.19%	72.30%	95.3%	95.30%	96.00%

Time:
2.51× speedup

Accuracy:
6% increase

Memory
< 15MB

Task	Memory Footprint (MB)				
	RS	HL	GN	ODE-Est	ODE-Simplified
IC	1.70	11.91	16.89	18.27	16.92
HAR	1.92	7.27	12.23	13.46	12.38
TC	0.75	10.58	19.65	25.15	14.47

Delay
1ms

Task	Evaluation Time (ms)				
IC	0.05	11.1	21.1	22.8	11.4
HAR	0.05	0.36	1.04	1.93	0.53
TC	0.05	1.03	9.06	9.69	1.23

Robustness Analysis

Number of Local Training Epochs (m)

With the increasing of local training epoch number, ODE achieves hig her final accuracy improvement

Device Participation Rate

With 10% participation rate, ODE achieves 2.57× training time speed up and 6.6% increase on inference accuracy.

On-Device Storage Capacity

ODE has stable performance with different device capacity, and can r educe up to 50% storage compared with baseline .

Conclusion

- Identify **two practical properties** of mobile devices and demonstrate the enormity
- **Theoretically** analyze impact of an individual local data sample on global model
- Design a collaborative data selection **framework** for FL to **simultaneously** improve convergence rate and final inference accuracy
- Achieve **remarkable** training speedup and accuracy improvement on **industrial** traffic classification task

Thanks for Watching !

Q & A

Please refer to our paper for more details !