CVPR JUNE 17-21, 2024 SEATTLE, WA

THU-AM-308

Towards Efficient Replay in Federated Incremental Learning

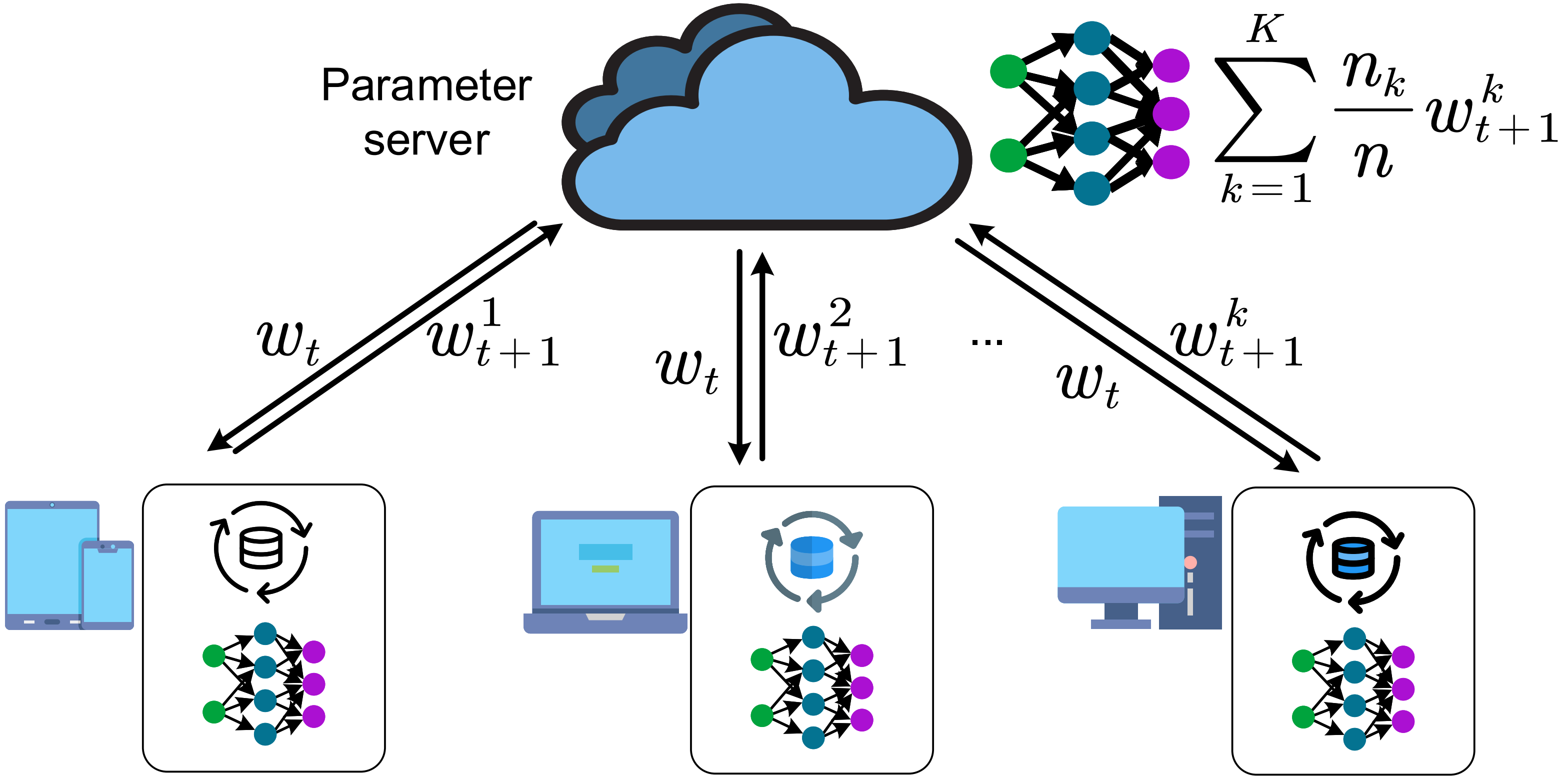
Yichen Li[1], Qunwei Li[2], Haozhao Wang[1], Ruixuan Li[1*], Wenliang Zhong[2], Guannan Zhang[2]

[1]Huazhong University of Science and Technology, Wuhan, China

[2]Ant Group, Hangzhou, China

Federated Incremental Learning

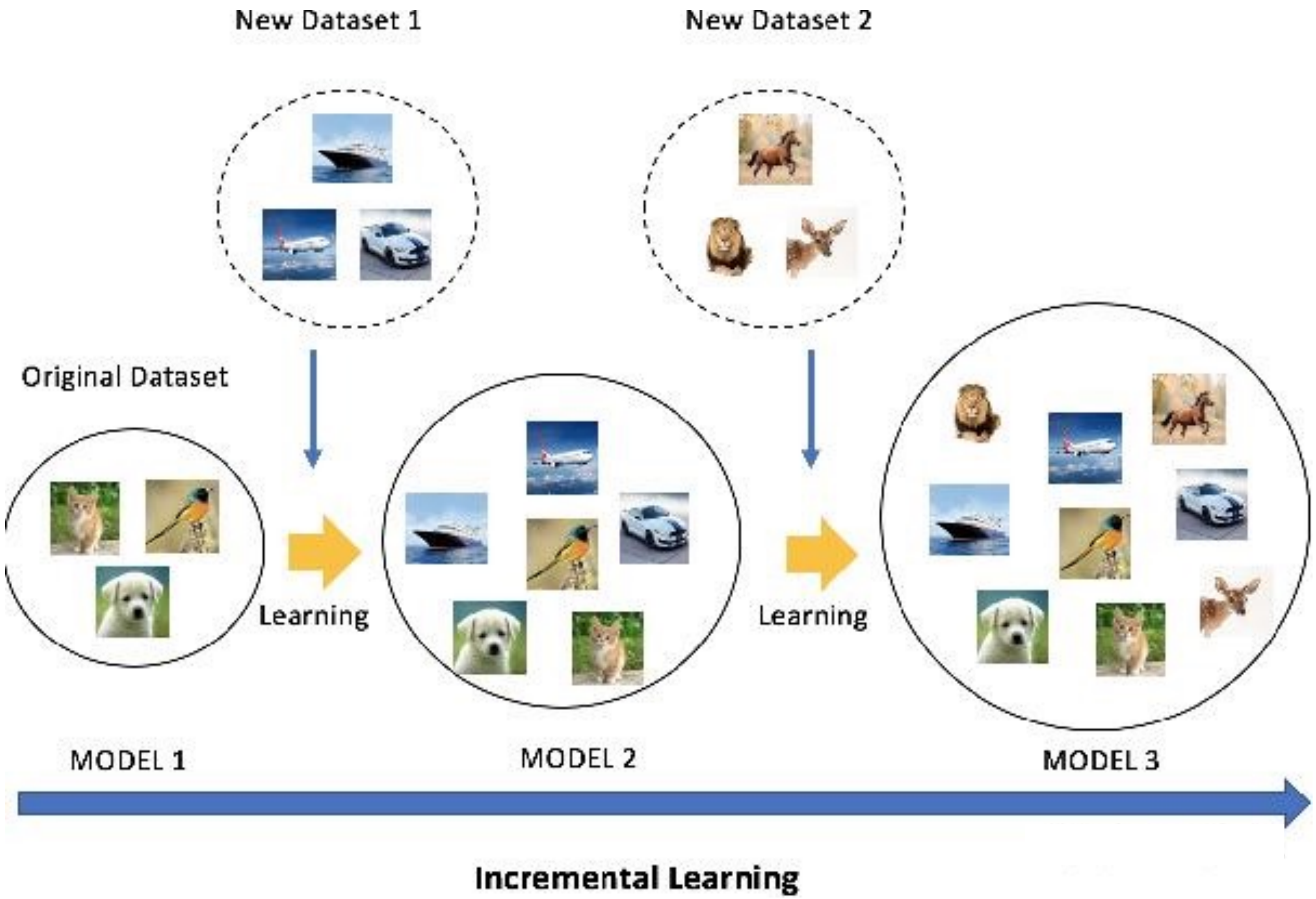
Background: Federated Learning



FedAvg: <u>Global model</u> is obtained by <u>computing the average</u> of <u>parameters</u> of multiple local models

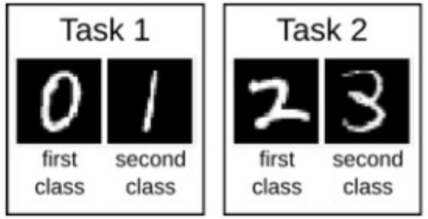
Background: Continual Learning

Illustration of Continual Learning/Incremental Learning/Lifelong Learning



Three Typical Scenarios

- **Class-Incremental Learning**



$$P(Y^1) \neq P(Y^2)$$

- **Domain-Incremental Learning**

$$P(X^1) \neq P(X^2)$$

- **Task-Incremental Learning**

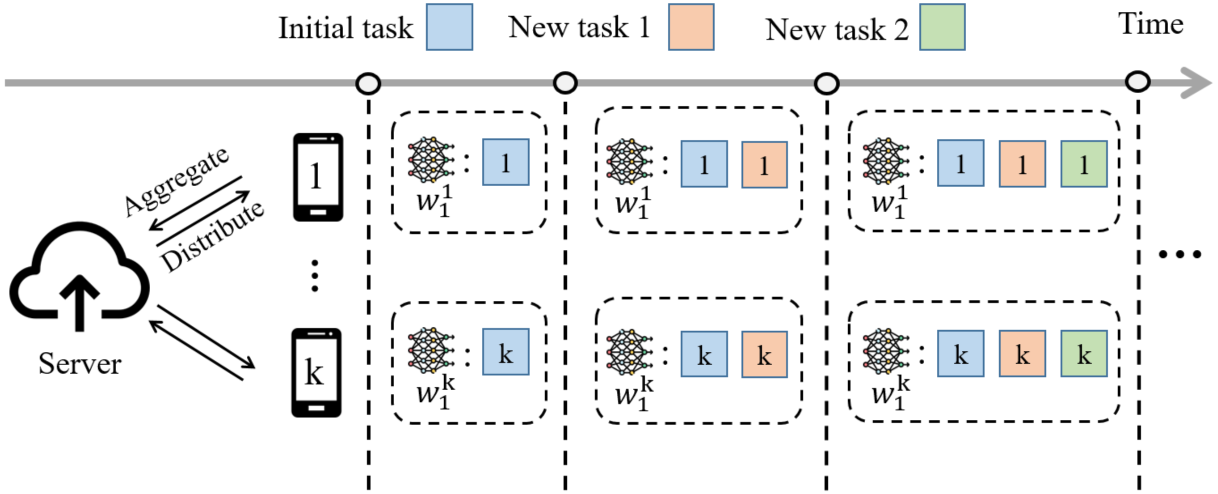
$$P(Y^1) \neq P(Y^2), P(X^1) \neq P(X^2), |Y^1| \neq |Y^2|$$

Limitations

- ◆ **Dynamic**: existing FL methods typically assume the data in each client is fixed or static.
- ➢ data often comes in an incremental manner, where the data domain may increase dynamically.

- ◆ **Catastrophic Forgetting**: clients are difficult to learn new data while retaining previous information
- ➢ especially when data is non-identically and independently distributed (Non-IID) across clients.

		Spatial heterogeneity			
		No changes (IID data)	Changes in the input space throughout clients $P_i(x) \neq P_j(x)$	Changes in the behaviour throughout clients $P_i(y\|x) \neq P_j(y\|x)$	Changes in the input space and behaviour throughout clients
Temporal heterogeneity	No changes (IID data)	OUT OF SCOPE	[35–38, 44] [55, 61–81]	[39–42] [99–101]	[45] [50–52]
	Changes in the input space over time, $P^t(x) \neq P^{t+k}(x)$ (Virtual Concept Drift)	[8, 117–122] [131–137]	NOT ADDRESSED SO FAR		
	Changes in the behaviour over time, $P^t(y\|x) \neq P^{t+k}(y\|x)$ (Real Concept Drift)	[123–125] [138–144]			
	Changes in the input space and behaviour over time (Total Concept Drift)	NO SPECIFIC ALGORITHMS			

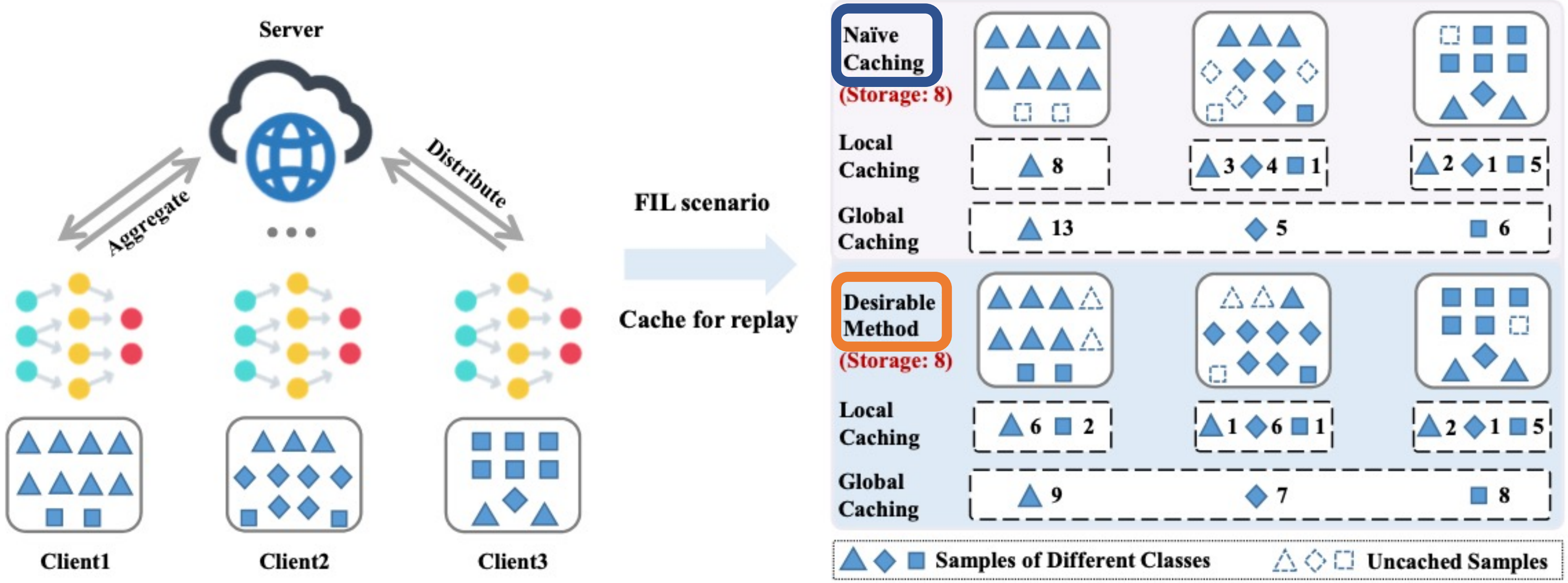


i.e. Domain-Incremental learning (DIL)

Motivations

- **Assumption**: each client can cache a few samples with the local storage for replay
  - lack enough storage space to retain full data

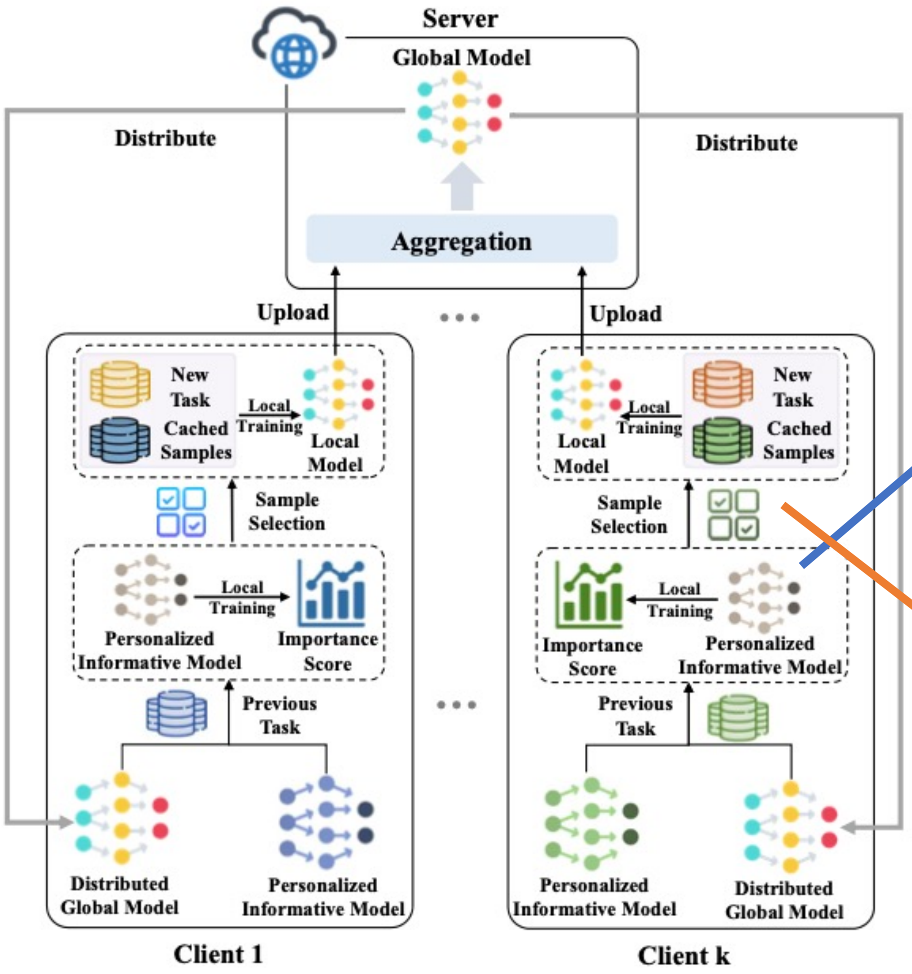
An example of 3-client in FIL scenario



Synergistic Replay with Important Samples !

Methodology: Re-Fed

- How to ensure that samples can balance local training and global data distribution?
- How to quantify the importance of samples?



◆ Personalized Informative Model

$$v_{k,s}^{t-1} = v_{k,s-1}^{t-1} - \eta\Big(\sum_{i=1}^{M} \nabla l\big(f_{v_{k,s-1}^{t-1}}(\tilde{x}_{k,t-1}^{(i)}), \tilde{y}_{k,t-1}^{(i)}\big) + q(\lambda)(v_{k,s-1}^{t-1} - w^{t-1})\Big). \qquad (3)$$

◆ Sample Importance Score

$$G^p(\tilde{x}_{k,t-1}^{(i)}) = \left\|\nabla l\left(f_{v_{k,p}^{t-1}}(\tilde{x}_{k,t-1}^{(i)}), \tilde{y}_{k,t-1}^{(i)}\right)\right\|^2. \qquad (4)$$

$$I(\tilde{x}_{k,t-1}^{(i)}) = \sum_{p=1}^{s} \frac{1}{p} G^p(\tilde{x}_{k,t-1}^{(i)}). \qquad (5)$$

Theory

Lemma 1 (Proportion of Global and Local Information.) *For all $\lambda \in (0,1)$ and $\lambda \to f(\lambda_k)$ is non-increasing:*

$$\frac{\partial \nabla f(\hat{v_k}(\lambda))}{\partial \lambda} \leq 0$$
$$\frac{\partial ||\hat{v_k}(\lambda) - \hat{w}||}{\partial \lambda} \geq 0. \tag{11}$$

Then, for $k \in [K]$, we can get:

$$\lim_{\lambda \to 0} \hat{v_k}(\lambda) := \hat{w}. \tag{12}$$

Lemma 2 ([21] Lemma 13.) Under assumptions above, $f(v_k)$ is μ_k-strongly convex at each communication round t, we have:

$$\mathbb{E}\left[||v_k^{t+1} - \hat{v_k}||^2\right] \leq (1 - \eta(\mu_k + q(\lambda)))\,\mathbb{E}\left[||v_k^t - \hat{v_k}||^2\right] + \eta^2\left(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k})\right)^2 + \eta^2 q(\lambda)^2 \mathbb{E}\left[||w^t - \hat{w}||^2\right]$$
$$+ 2\eta^2 q(\lambda)\left(\sigma + q(\lambda)(M + \frac{\sigma}{\mu_k})\right)\sqrt{\mathbb{E}\left[||w^t - \hat{w}||^2\right]} + 2\eta q(\lambda)\sqrt{\mathbb{E}\left[||v_k^t - \hat{v_k}||^2\right]\mathbb{E}\left[||w^t - \hat{w}||^2\right]}. \tag{13}$$

Theorem 3.1 (Personalized Informative Model.) *Assuming the global model w^t converges to the optimal model $\hat{w}$ with $g(t)$ for any client $k \in [K]$ at each communication round t: $\mathbb{E}\left[||w^t - \hat{w}||^2\right] \leq g(t)$ and $\lim_{t\to\infty} g(t) = 0$, then there exists a constant $C < \infty$ such that the personalized informative model v_k^t can converge to the optimal model $\hat{v_k}$ with $Cg(t)$.*

Experiments - Settings

Datasets

Class-Incremental Learning

- CIFAR10
- CIFAR100
- Tiny-ImageNet

Domain-Incremental Learning

- Digit10
- Office31
- Domain Net

Baselines

Traditional FL Methods

- FedAvg
- FedProx

Customed Methods

- Fixed
- DANN+FL
- Shared

Existing FIL Methods

- FCIL
- FedDIL

Experiments - Performance Overview

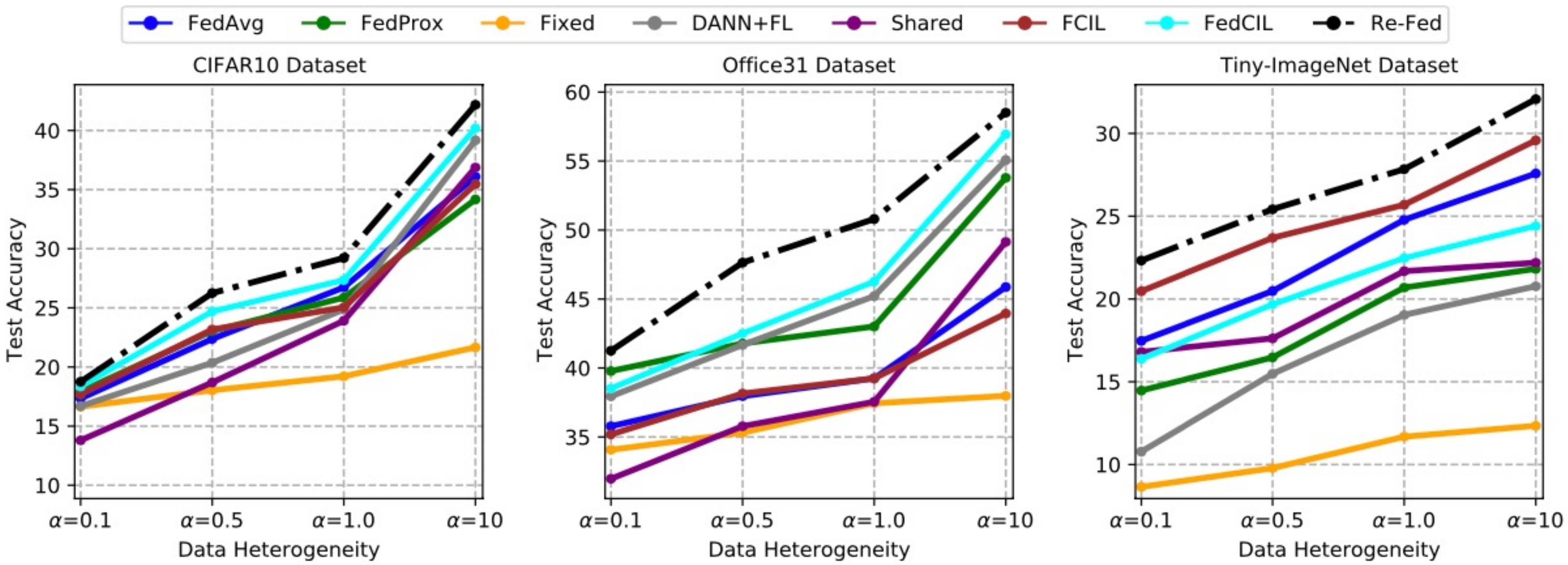
Test Accuracy & Communication Efficiency

Scenario	Dataset	FedAvg	FedProx	Fixed	DANN+FL	Shared	FCIL	FedCIL	Re-Fed
Class-Incremental	CIFAR10 ($\alpha = 1.0$)	26.73±1.12	25.87±0.68	19.21±0.06	24.86±2.31	23.91±1.70	25.04±0.11	27.35±1.24	**29.22±0.49**
	CIFAR100 ($\alpha = 5.0$)	17.21±1.35	18.03±0.91	9.27±0.22	19.73±2.17	18.30±1.53	23.02±0.66	17.98±1.46	**25.61±0.88**
	Tiny-ImageNet ($\alpha = 10$)	27.58±0.74	21.82±0.90	12.34±0.23	20.77±1.31	22.19±0.54	29.58±0.15	24.41±0.95	**32.07±0.27**
Domain-Incremental	Digit10 ($\alpha = 0.1$)	77.59±0.39	79.09±0.58	71.26±0.04	76.44±1.05	74.77±0.23	77.59±0.39	83.85±0.80	**85.96±0.14**
	Office31 ($\alpha = 1$)	39.25±1.61	43.01±1.59	37.44±0.72	45.21±2.10	37.55±0.69	39.25±1.61	46.26±2.24	**50.80±0.77**
	DomainNet ($\alpha = 10$)	51.73±2.32	49.12±2.71	46.30±1.42	50.01±3.31	41.76±1.26	51.73±2.32	47.28±3.01	**56.66±0.50**

Scenario	Dataset	FedAvg	FedProx	Fixed	DANN+FL	Shared	FCIL	FedCIL	Re-Fed
Class-Incremental	CIFAR10 (Task:5)	613±2.67	685±3.00	142±0.67	712±3.67	574±1.33	590±2.67	738±4.00	**562±1.67**
	CIFAR100 (Task:10)	1103±2.33	1246±3.00	137±2.00	1258±4.67	1154±3.33	1095±2.67	1311±5.67	**1039±4.33**
	Tiny-ImageNet (Task:10)	1197±2.67	1234±2.67	132±3.00	1305±3.67	1278±4.33	1185±2.33	1317±3.33	**1128±3.67**
Domain-Incremental	Digit10 (Task:4)	410±1.67	412±0.67	112±0.33	483±1.33	372±2.00	410±1.67	419±2.67	**325±1.33**
	Office31 (Task:3)	413±2.67	429±2.00	144±0.67	436±3.67	391±1.12	413±2.67	431±3.33	**388±1.67**
	DomainNet (Task:6)	726±3.33	767±2.67	141±1.67	752±4.00	694±2.67	726±3.33	791±3.67	**661±2.33**

Experiments - Performance Overview

Data Heterogeneity



a smaller α indicates higher data heterogeneity

$$v_{k,s}^{t-1} = v_{k,s-1}^{t-1} - \eta\Big(\sum_{i=1}^{M} \nabla l\left(f_{v_{k,s-1}^{t-1}}(\tilde{x}_{k,t-1}^{(i)}), \tilde{y}_{k,t-1}^{(i)}\right) + q(\lambda)(v_{k,s-1}^{t-1} - w^{t-1})\Big). \quad (3)$$

Dataset	$\alpha=0.1$			$\alpha=0.5$			$\alpha=1.0$			$\alpha=10$		
	$\lambda=0.2$	$\lambda=0.5$	$\lambda=0.8$	$\lambda=0.2$	$\lambda=0.5$	$\lambda=0.8$	$\lambda=0.2$	$\lambda=0.5$	$\lambda=0.8$	$\lambda=0.2$	$\lambda=0.5$	$\lambda=0.8$
CIFAR10	**18.75±1.30**	18.61±1.09	17.91±0.81	**26.25±1.64**	26.00±0.97	25.62±0.52	27.05±0.88	27.80±0.21	**29.22±0.49**	38.43±0.43	40.04±0.19	**42.17±0.25**
Office31	**41.25±1.01**	39.29±1.34	38.18±0.68	46.86±0.91	**47.64±0.53**	47.13±1.16	43.81±0.73	48.67±0.99	**50.08±0.77**	52.79±1.28	55.92±0.38	**58.51±0.46**
Tiny-ImageNet	**22.32±0.12**	20.51±0.98	18.00±1.30	24.60±0.48	**25.42±0.59**	24.39±0.66	24.88±0.87	27.15±0.78	**27.84±0.73**	29.03±0.30	30.26±0.24	**32.07±0.27**

Conclusions

We propose a simple framework called **Re-Fed** to address the issues of catastrophic forgetting and data heterogeneity in federated continual learning. It has the following advantages:

- ✓ **Optimization:** Re-Fed allows for the use of aggregation methods other than FedAvg to update the global model while maintaining convergence properties.
- ✓ **Privacy:** Unlike typical FL algorithms, Re-Fed does not transmit additional information over the network, thus avoiding privacy issues that arise from applying sample reconstruction methods for data replay.
- ✓ **Resources:** Re-Fed enables each client to train a base model using only its local training data without requiring additional distilled or generated augmented data, thereby avoiding extra computational costs or storage overhead.

Thank You

Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li,
Wenliang Zhong, Guannan Zhang