Fed-CO2: Cooperation of Online and Offline Models for Severe Data Heterogeneity in Federated Learning



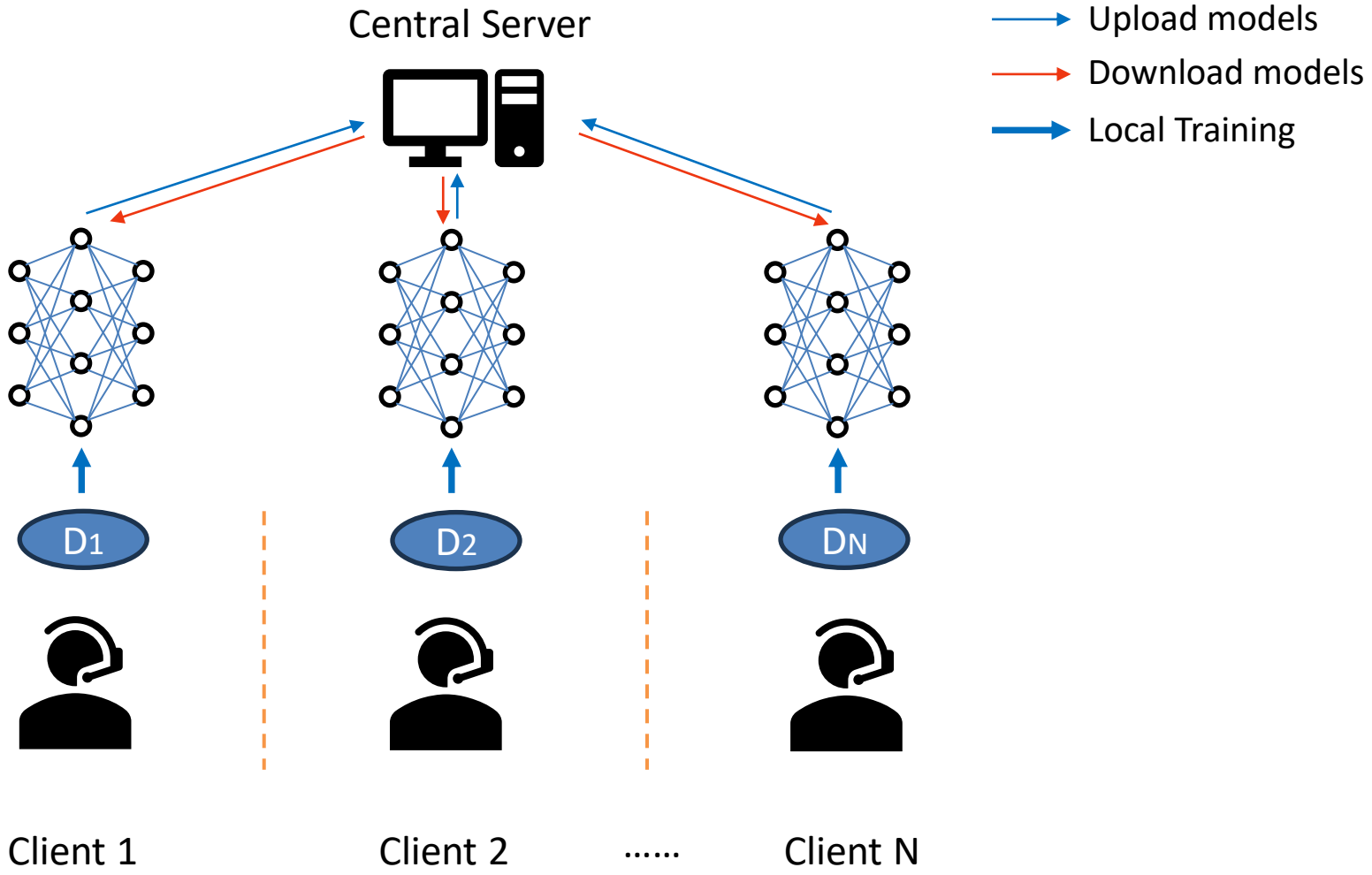

Zhongyi Cai[1] Ye Shi[1]* Wei Huang[2] Jingya Wang[1]

[1]ShanghaiTech University [2]RIKEN Center for Advanced Intelligence Project

NeurIPS 2023
October 27, 2023

Federated Learning

·What is Federated Learning?
For Collaboration and Privacy!



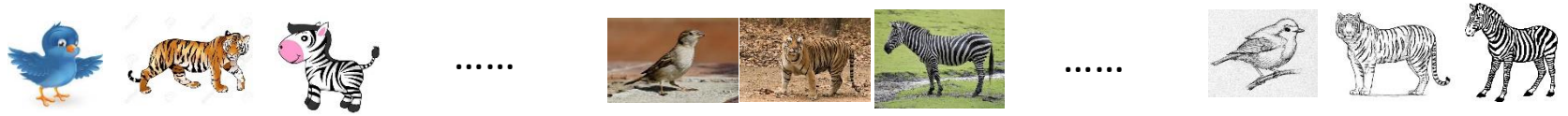
Motivation

·Design a universal framework to address diverse data heterogeneity issues!

Most previous studies focused on label distribution imbalance data heterogeneity issues

Client 1 Client k Client N

Few works researched feature shift data heterogeneity issues



Client 1 Client k Client N

So far, these two forms of data heterogeneity have been studied separately and have not been addressed within a unified FL framework.

Observation

·Several existing algorithms personalized some parts of the model in FL with label distribution imbalance or feature shift to retain and leverage some of the local offline information.

·However, in realistic scenarios, where extreme label distribution skew, feature skew, or even both are present, these algorithms fail to effectively harness local specialized knowledge for satisfactory adaptation.

·Especially, in some cases of extreme heterogeneity, models trained by existing personalized algorithms may even perform worse than the locally trained model.

·On the other hand, in FL with milder heterogeneity, partially personalized models perform better due to their ability to access online general information from other clients.

Therefore, here comes the question:

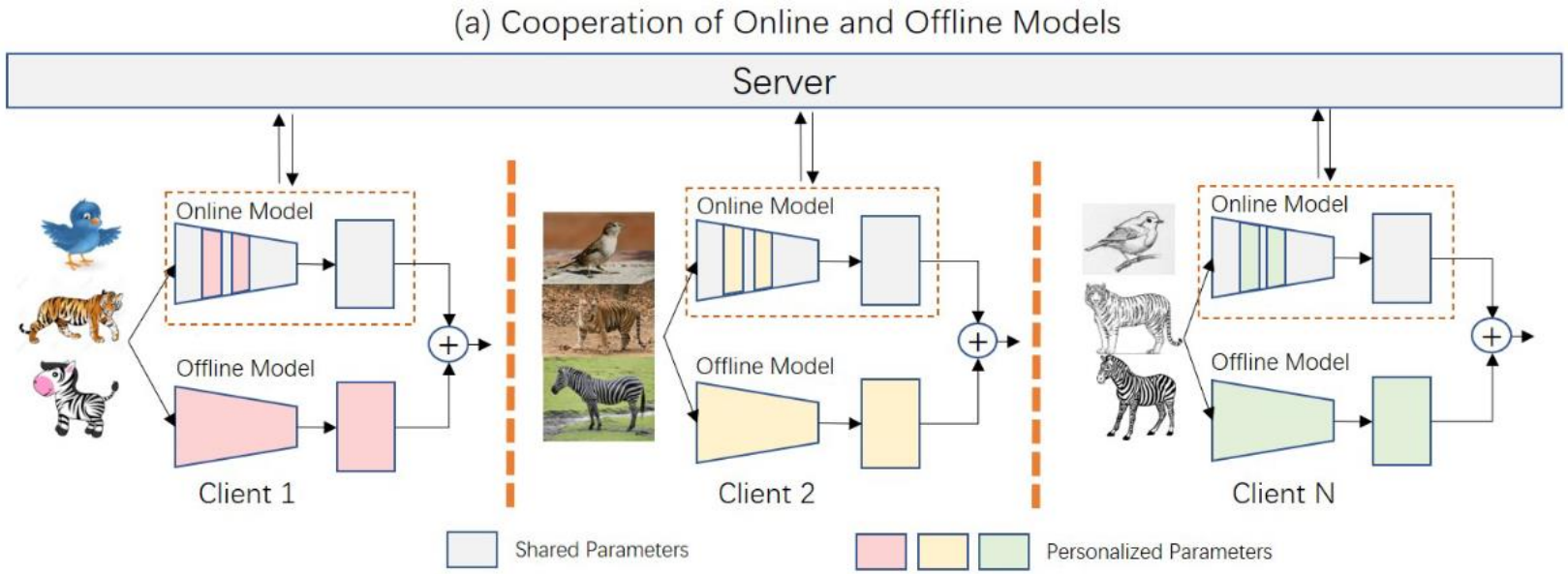
Is there a more effective approach to fuse the online general knowledge and the offline specialized knowledge for better performance?

Algorithm Framework

·Universal Federated Learning Framework

Online Model: Partially Personalized Model learns general knowledge.
Offline Model: Locally Trained Model learns specialized knowledge.
With prediction fusion, Fed-CO2 combines general knowledge and specialized knowledge.

$$F_i(\theta_i\,;\,x) = F_i^{\text{on}}(\theta_{p,i}^{\text{on}}\,,\,\theta_g^{\text{on}}\,;\,x) + F_i^{\text{off}}(\theta_i^{\text{off}}\,;\,x).$$



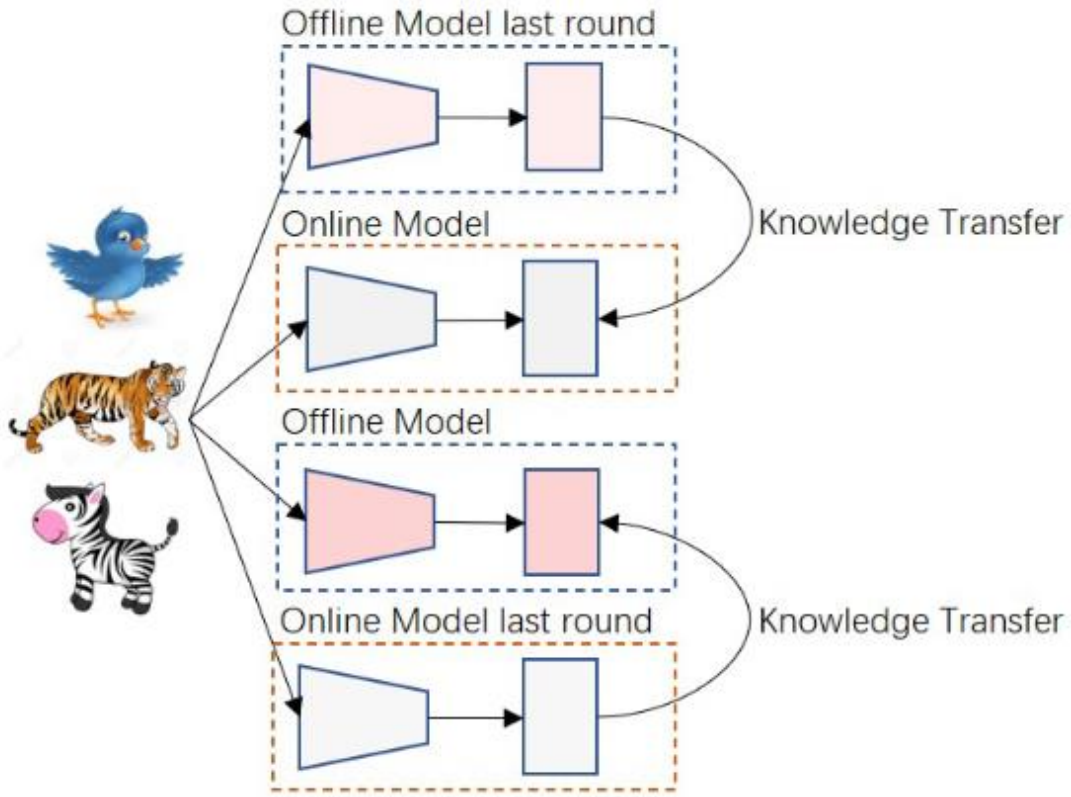
Intra-Client Knowledge Transfer

·Model-level Cooperation

The online model learns local knowledge and the offline model learns global knowledge.

$$\theta_{i,s}^{\text{on}} = \theta_{i,s-1}^{\text{on}} - \alpha \cdot \nabla_{\theta_{i,s-1}^{\text{on}}} KL(\bar{F}_i^{\text{off}}(\bar{\theta}_i^{\text{off}}\,;\, x_k)\,,\, F_i^{\text{on}}(\theta_{i,s-1}^{\text{on}}\,;\, x_k)),$$

$$\theta_{i,s}^{\text{off}} = \theta_{i,s-1}^{\text{off}} - \alpha \cdot \nabla_{\theta_{i,s-1}^{\text{off}}} KL(\bar{F}_i^{\text{on}}(\bar{\theta}_i^{\text{on}}\,;\, x_k)\,,\, F_i^{\text{off}}(\theta_{i,s-1}^{\text{off}}\,;\, x_k)),$$



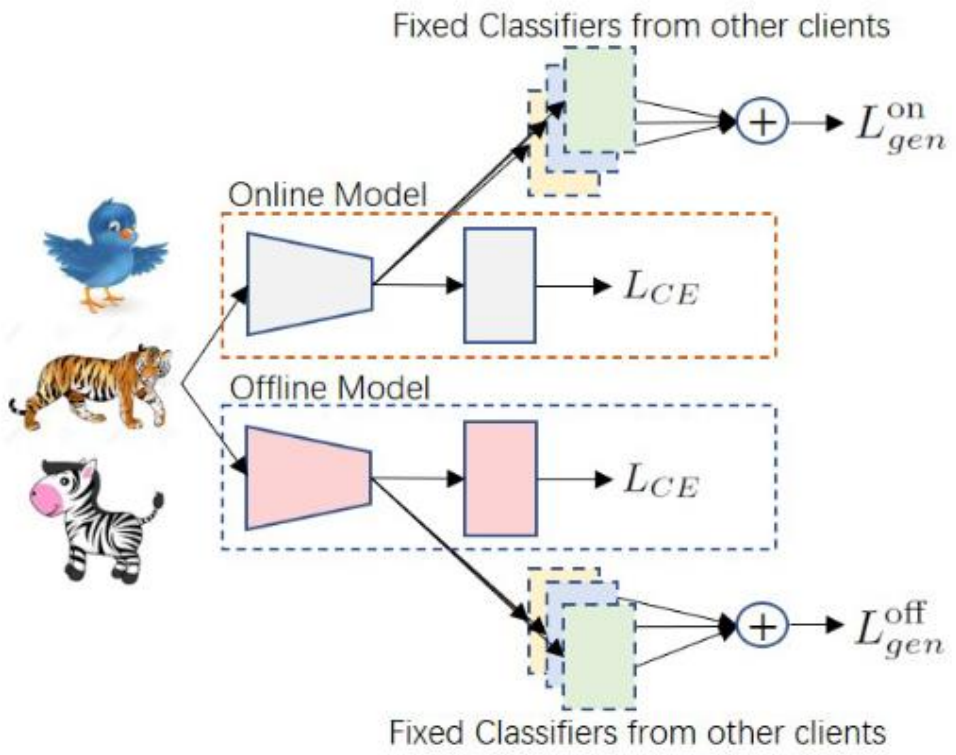
Inter-Client Knowledge Transfer

·Client-level Cooperation

Leverage additional knowledge from other clients to enhance models' domain generalization ability.
Utilize global domain-invariant knowledge to benefit local task.

$$L_{gen}\left(x_k\,,\,y_k\right)=\sum\nolimits_{j\neq i}L_{CE}\left(\overline{C}_j^{\text{off}}\left(\overline{\phi}_j^{\text{off}};\,f_i\left(\eta_{i,t};\,x_k\right)\right),\,y_k\right),$$

$$L=L_{CE}(x_k\,,\,y_k)+\mu\cdot L_{gen}(x_k\,,\,y_k),$$



Theoretical Analysis

·Convergence Proof with the Theory Neural Tangent Kernel (NTK)

Following previous work[1] , we can decompose the NTK in a direction component and a magnitude component.

$$\frac{d\mathbf{F}}{dt} = -\mathbf{\Lambda}(t)(\mathbf{F}(t) - \mathbf{y}), \mathbf{\Lambda}(t) := \frac{\mathbf{V}(t)}{\alpha^2} + \mathbf{G}(t).$$

When α ≤ 1, the convergence rate is dominated by **V**(t).
In this case, the convergence performance can be analyzed by comparing $\lambda_{\min}(\mathbf{V}^{\infty})$.

Theorem 4.3 *For the V-dominated convergence, the convergence rate of Fed-CO_2 is faster than that of FedBN (the online model in Fed-CO_2).*

Objective: $\lambda_{\min}(\mathbf{V}^{\infty}) \geq \lambda_{\min}(\mathbf{V}_{\text{on}}^{\infty})$.

Proof Sketch:
Step1: Demonstrate the offline model converges faster than FedBN by proving $\lambda_{\min}(\mathbf{V}_{\text{off}}^{\infty}) \geq \lambda_{\min}(\mathbf{V}_{\text{on}}^{\infty})$.

Step2: Given that $\lambda_{\min}(\mathbf{V}^{\infty}) = \lambda_{\min}(\frac{1}{2}(\mathbf{V}_{\text{on}}^{\infty} + \mathbf{V}_{\text{off}}^{\infty})) \geq (\frac{1}{2}\lambda_{\min}(\mathbf{V}_{\text{on}}^{\infty}) + \frac{1}{2}\lambda_{\min}(\mathbf{V}_{\text{off}}^{\infty}))$, we can prove that $\lambda_{\min}(\mathbf{V}^{\infty}) \geq \lambda_{\min}(\mathbf{V}_{\text{on}}^{\infty})$.

[1]Optimization theory for relu neural networks trained with normalization layers (ICML.2020)

Empirical Results

·FL with Feature Skew

Table 1: Experiment results for FL with Feature Skew on Office-Caltech10.

Methods	Office-Caltech10				
	Amazon	Caltech	DSLR	WebCam	Avg
SingleSet	54.9±1.5	40.2±1.6	78.7±1.3	86.4±2.4	65.1±1.7
FedAvg [2]	54.1±1.1	44.8±1.0	66.9±1.5	85.1±2.9	62.7±1.6
FedProx [4]	54.2±2.5	44.5±0.5	65.0±3.6	84.4±1.7	62.0±2.1
FedPer [11]	49.0±1.2	37.1±2.4	57.7±3.7	79.7±2.1	56.0±1.1
MOON [7]	57.3±0.7	44.4±0.5	76.2±2.5	83.1±1.1	65.2±0.5
FedRod [12]	60.4±2.3	45.3±0.9	73.7±2.5	83.7±2.3	65.8±1.4
COPA [33]	51.9±2.5	46.7±0.8	65.6±2.0	85.0±1.3	62.3±0.9
FedBN [10]	63.0±1.6	45.3±1.5	83.1±2.5	90.5±2.3	70.5±2.0
Fed-CO_2	**63.0±1.6**	**49.1±0.7**	**89.4±2.5**	**96.6±1.5**	**74.5±0.3**

Table 2: Experiment results for FL with Feature Skew on DomainNet.

Methods	DomainNet						
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
SingleSet	41.0±0.9	23.8±1.2	36.2±2.7	73.1±0.9	48.5±1.9	34.0±1.1	42.8±1.5
FedAvg [2]	48.8±1.9	24.9±0.7	36.5±1.1	56.1±1.6	46.3±1.4	36.6±2.5	41.5±1.5
FedProx [4]	48.9±0.8	24.9±1.0	36.6±1.8	54.4±3.1	47.8±0.8	36.9±2.1	41.6±1.6
FedPer [11]	40.4±0.8	25.7±0.6	37.3±0.6	62.5±1.2	47.4±0.5	32.8±0.8	41.0±0.3
MOON [7]	52.5±1.1	25.7±0.6	39.4±1.7	50.8±4.7	48.8±0.8	40.1±4.1	42.9±1.5
FedRoD [12]	50.8±1.6	26.3±0.2	40.1±1.8	66.8±1.8	51.5±1.1	39.1±2.0	45.7±0.7
COPA [33]	51.1±1.0	24.7±1.2	36.8±0.8	54.8±1.6	47.1±1.8	41.0±1.4	42.6±0.4
FedBN [10]	51.2±1.4	26.8±0.5	41.5±1.4	71.3±0.7	54.8±0.8	42.1±1.3	48.0±1.0
Fed-CO_2	**55.0±1.1**	**28.6±1.1**	**44.3±0.6**	**75.1±0.6**	**62.4±0.8**	**45.7±1.9**	**51.8±0.2**

Empirical Results

·FL with label distribution skew

Table 3: Experiment results for FL with Label Distribution Skew on CIFAR10 and CIFAR100. Experiments are conducted with two kinds of label distribution data heterogeneity: Pathological setting and Dirichlet setting.

Methods	CIFAR10		CIFAR100	
	Pathological	Dirichlet	Pathological	Dirichlet
SingleSet	85.85±0.05	68.38±0.06	49.54±0.05	21.39±0.05
FedAvg [2]	44.12±3.10	57.52±1.01	14.59±0.40	20.34±1.34
FedProx [4]	57.38±1.08	56.46±0.66	21.32±0.71	19.40±1.76
FedPer [11]	80.99±0.71	74.21±0.07	42.08±0.18	20.06±0.26
MOON [7]	48.43±3.18	54.49±1.87	17.89±0.76	19.73±0.71
FedRoD [12]	**89.05±0.04**	73.99±0.09	54.96±1.30	28.29±1.53
FedBN [10]	86.71±0.56	75.41±0.37	48.37±0.56	28.70±0.46
Fed-CO_2	88.79±0.25	**77.45±0.30**	**58.50±0.43**	**32.43±0.37**

·FL with label distribution skew and feature skew

Table 4: Experiment results for FL with both Label Skew and Feature Skew on Digits.

Methods	Digits					
	MNIST	SVHN	USPS	SynthDigits	MNIST-M	Avg
SingleSet	83.75±7.58	74.63±0.31	97.14±0.06	87.95±0.46	80.55±0.26	84.80±1.44
FedAvg [2]	89.27±6.39	57.23±5.43	94.60±1.05	81.30±2.51	71.71±7.59	78.82±4.03
FedProx [4]	87.09±7.83	53.40±7.38	90.55±7.47	78.40±7.05	69.98±6.55	75.88±6.90
FedPer [11]	96.92±0.02	72.86±0.11	97.09±0.12	87.82±0.02	83.69±0.07	87.68±0.02
MOON [7]	96.58±0.07	74.10±0.22	96.19±0.10	88.15±0.04	85.05±0.14	88.01±0.07
FedRod [12]	96.65±0.06	77.05±0.16	96.88±0.13	89.59±0.06	86.18±0.11	89.27±0.05
COPA [33]	96.82±0.08	78.32±0.12	96.54±0.15	89.36±0.04	87.46±0.11	89.70±0.06
FedBN [10]	92.68±3.45	70.26±4.38	91.40±8.72	83.17±4.95	77.98±3.84	83.10±4.96
Fed-CO_2	**97.17±0.67**	**83.16±0.23**	**98.12±0.13**	**93.04±0.11**	**91.45±0.14**	**92.59±0.15**

Conclusion

We propose a universal Federated Learning framework that could handle both label distribution skew and feature skew data heterogeneity issues within a **C**ooperation mechanism between the **O**nline and **O**ffline models, namely **Fed-CO2**.

·Unified cooperation FL framework which utilizes both local specialized knowledge and global general knowledge.

·Enhanced cooperation mechanisms in both model-level (Intra-client) and client-level (Inter-client) for more severe feature shift skew issues.

·Provide solid empirical and theoretical results.

Codes have been published at: https://github.com/zhyczy/Fed-CO2.