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

*Corresponding Author

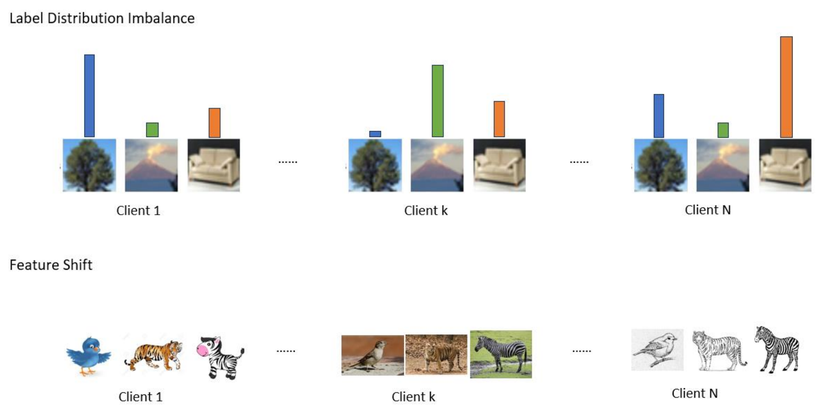
NEURAL INFORMATION PROCESSING SYSTEMS

Background

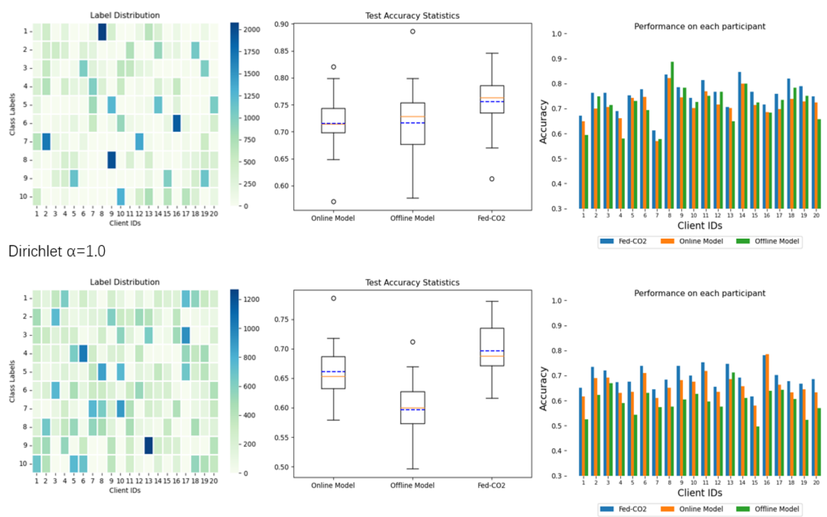
- Federated Learning is a collaborative learning paradigm, where multiple clients collaboratively learn a global model without sharing their private data.
- When data among clients are not independently and identically distributed, the performance of the learned consensus model can degrade substantially. This problem is called the data heterogeneity issue.

Motivation

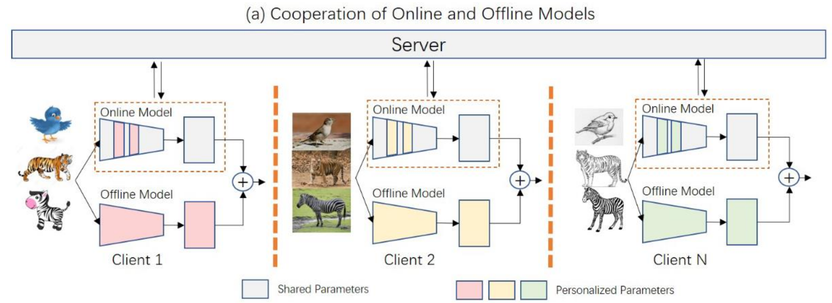
Address both label distribution imbalance and feature shift data heterogeneity issues in Federated Learning.



Make full use of local specialized knowledge and global general knowledge.



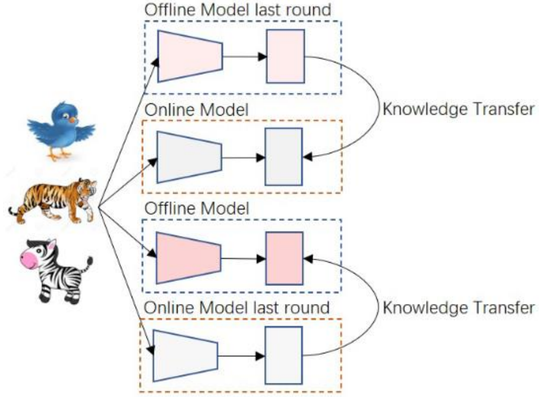
Universal Framework



- A universal Federated Learning framework for both label distribution skew and feature skew with the cooperation of online and offline models.
- Online model is partially personalized and learns general knowledge.
- Offline model is locally trained and learns specialized knowledge.
- In the test phase, we fuse predictions from online and offline models to combine general knowledge and specialized knowledge.

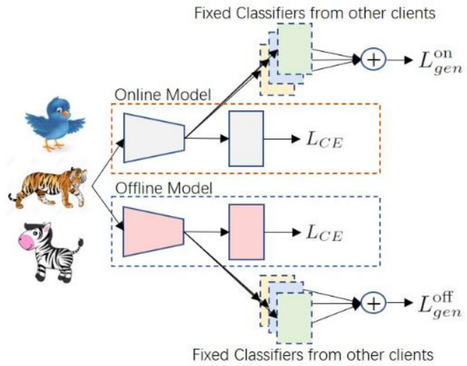
Enhanced Cooperation Mechanisms

Intra-Client Knowledge Transfer Mechanism.



- Utilize knowledge distillation with KL divergence to transfer knowledge between the online and offline models.
- Freeze the teacher models to stabilize learned knowledge.

Inter-Client Knowledge Transfer Mechanism.



- Introduce classifiers of the offline model from other clients to access general knowledge from other clients.
- Let image features be recognized well by these introduced classifiers to enhance model's domain generalization ability.

Convergence Analysis

We theoretically demonstrate Fed-CO2 converges faster than FedBN[1].

Following previous work[2], 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 $\alpha \leq 1$, the convergence rate is dominated by V(t).

In this case, the convergence performance can be analyzed by comparing the least eigenvalue of $\mathbf{V}^\infty, \lambda_{\min}(\mathbf{V}^\infty)$.

Theorem: For the V-dominated convergence, the convergence rate of Fed-CO2 is faster than that of FedBN.

Empirical Results

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 [34]	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 [34]	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**

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**

We achieve SOTA performance on FL scenarios with various kinds and degrees of data heterogeneity issues!

1. Li, Xiaoxiao, et al. "Fedbn: Federated learning on non-iid features via local batch normalization." ICLR (2021).
2. Dukler, Yonatan, Quanquan Gu, and Guido Montúfar. "Optimization theory for relu neural networks trained with normalization layers." International conference on machine learning. PMLR, 2020.