Class-Wise Federated Learning for Efficient Personalization

Gyuejeong Lee, Daeyoung Choi

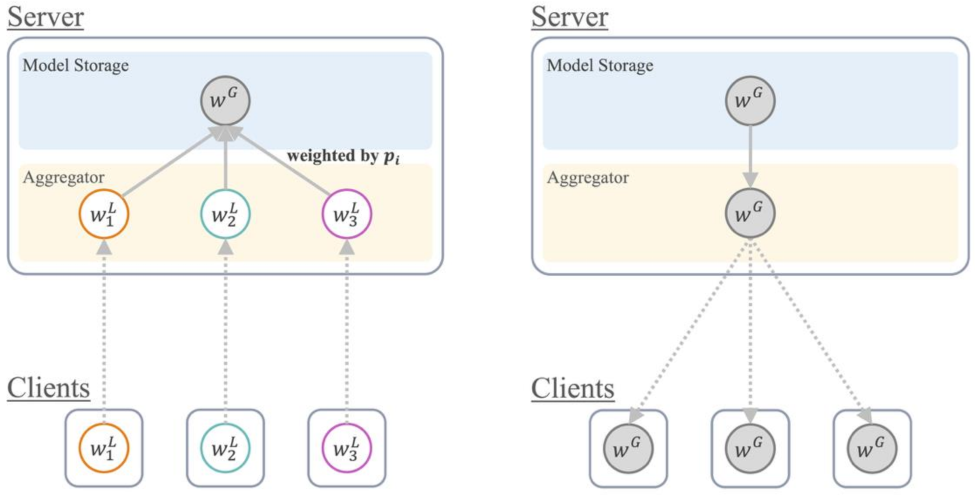
cwFedAvg | Background

A Foundational Approach of Federated Learning performs poorly with heterogeneous data distributions

- Federated learning(FL) enables distributed training without centralizing data, addressing challenges related to data costs and privacy constraints.
- With non-independent and identically distributed (non-IID) data, FedAvg suffers performance degradation due to its lack of personalization capability.

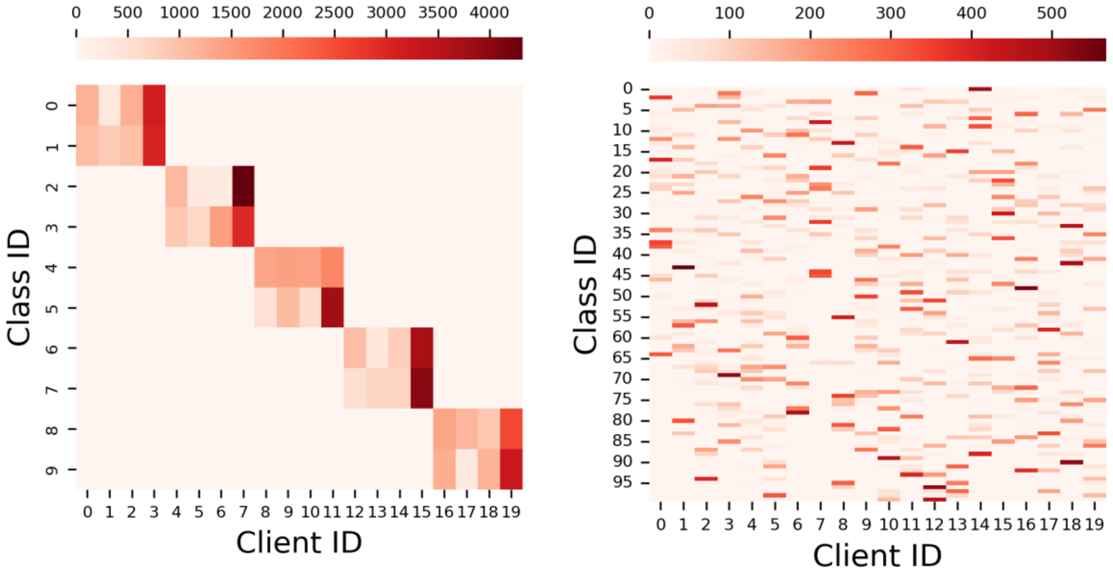
Standard Federated Learning Procedure(FedAvg)

- Federated Averaging(FedAvg) creates a single global model by aggregating local models weighted by client sample sizes.



<Clients → Server>   <Server → Clients>

The Heatmaps of non-IID Data Examples



<CIFAR-10 Pathological Setting>   <CIFAR-100 Practical Setting(0.1)>

cwFedAvg | Motivation

Why does FedAvg Fail to Train Personalized Models? It Fails to Capture Class-Specific PATHWAYs

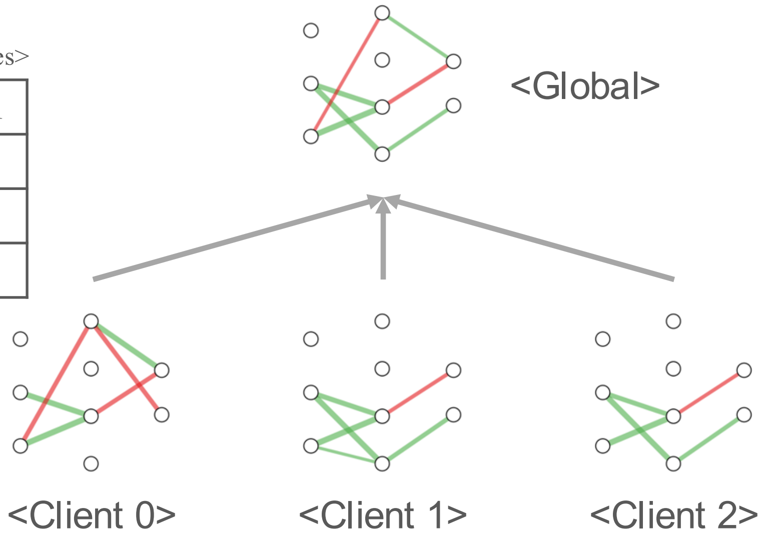
- Deep networks develop PATHWAYs, where a pathway represents a collection of paths (weights) connecting input to output.
- Pathways demonstrate distinct patterns across different classes based on the class proportion.

1. Class-Specific Pathway Collapse

- The learned pathways of local models show distinctive patterns.
- However, the global model cannot capture the unique patterns of each client.

<Synthetic non-IID Data Samples>

Client	Class 0	Class 1
0	2700	300
1	200	1800
2	500	500



2. Correlation btw weight vector and class distribution

- Based on the the correlation between the squared L_2-norms of the gradients of weight vectors and the number of samples of clients,
- We found L_2-norms of weight vectors of output layer correlate with the client's class distribution.

Theorem from [1]

$$\frac{\mathbb{E}\,\|\nabla\mathcal{L}(\mathbf{w}_{i,j})\|_2^2}{\mathbb{E}\,\|\nabla\mathcal{L}(\mathbf{w}_{i,k})\|_2^2} \approx \frac{n_{i,j}^2}{n_{i,k}^2} \Rightarrow \tilde{p}_{i,j} = \frac{\|\mathbf{w}_{i,j}\|_2}{\sum_{k=1}^{K}\|\mathbf{w}_{i,k}\|_2}$$

$\boldsymbol{w}_{i,j}$: the weight vector from penultimate layer neurons to output neuron j of client i
$n_{i,j}$: the number of samples belonging to class j of client i
$p_{i,j}$: the empirical class distribution of class j of client i
$\tilde{p}_{i,j}$: the approximated class distribution of class j of client i
K: the total number of class

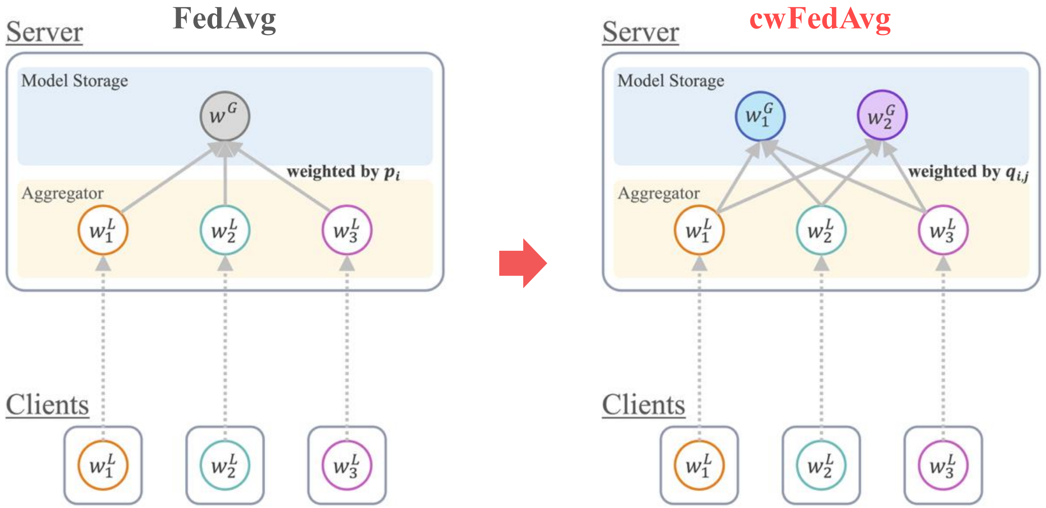
[1] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. IEEE transactions on neural networks, 4(6):962–969, 1993.

cwFedAvg | Method

Class-Wise Extension of FedAvg with Multiple Global Models and Modified Weighting Factors

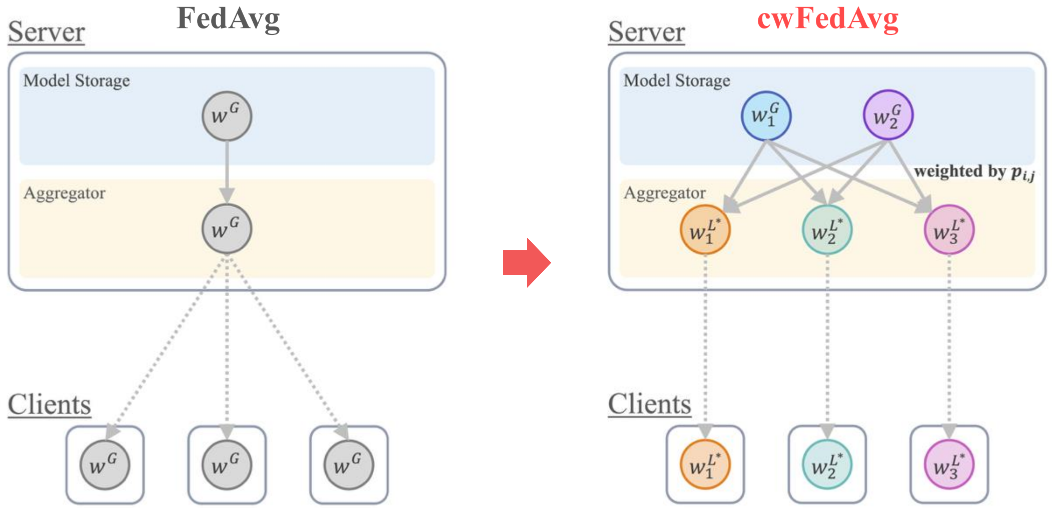
- Create class-specific global models by aggregating local models weighted by their respective client and class sample proportions.
- Generate personalized local models by aggregating these class-specific global models weighted by the class distribution of each client.

Clients → Server



$$\boldsymbol{w}_j^G = \sum_{i=1}^{M} q_{i,j} \boldsymbol{w}_i^L = \sum_{i=1}^{M} \frac{p_i \cdot \tilde{p}_{i,j}}{\sum_{i=1}^{M} p_i \cdot \tilde{p}_{i,j}} \boldsymbol{w}_i^L$$

Server → Clients



$$\boldsymbol{w}_i^L = \sum_{j=1}^{K} \tilde{p}_{i,j} \boldsymbol{w}_j^G$$

cwFedAvg | Method

WDR Enhances the Correlation btw Class Distribution and Model Weights, Improving Effectiveness

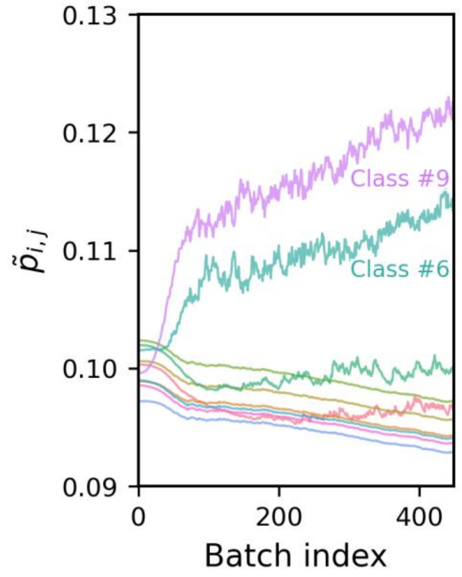
Weight Distribution Regularizer(WDR)

$$\mathcal{R}_i = \|\boldsymbol{p}_i - \tilde{\boldsymbol{p}}_i\|_2$$

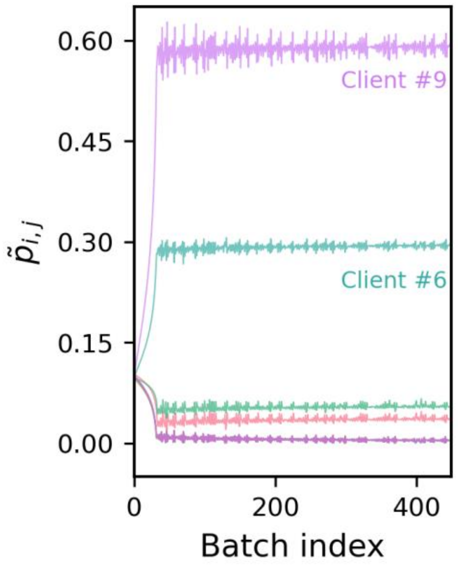
Total Cost Function

$$\tilde{\mathcal{L}}_i = \mathcal{L}_i + \lambda \mathcal{R}_i$$

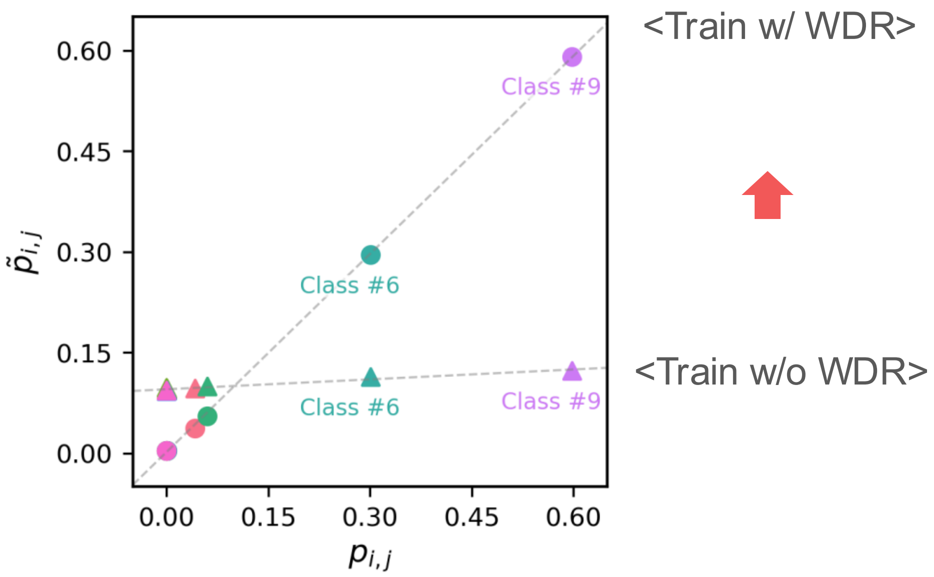
p_i: the empirical class distribution of client i
$\tilde{p}_i$: the approximated class distribution of client i
$\mathcal{L}_i$: the cost function
λ: the regularization coefficient



<Train w/o WDR>



<Train w/ WDR>



<Correlation>

cwFedAvg | Experiments(Quantitative)

cwFedAvg Consistently Outperforms All Other Settings Except for CIFAR-10 Practical Setting

Table 1. Classification accuracy (%) across datasets. Tiny ImageNet* indicates experiments using ResNet-18. `cwFedAvg` (Output) denotes `cwFedAvg` selectively applied to the output layer.

Algorithm	Pathological setting		Practical setting ($\alpha = 0.1$)				
	CIFAR-10	CIFAR-100	MNIST	CIFAR-10	CIFAR-100	Tiny ImageNet	Tiny ImageNet*
FedAvg	60.68±0.84	28.22±0.32	98.70±0.04	61.94±0.56	32.44±0.42	21.35±0.12	24.71±0.15
FedProx	60.65±0.92	28.59±0.28	98.68±0.09	62.48±0.86	32.26±0.26	20.65±0.12	24.06±0.16
FedAMP	88.82±0.15	63.29±0.49	99.26±0.01	**89.46±0.11**	47.65±0.62	29.95±0.10	31.38±0.18
FedFomo	90.76±0.59	63.12±0.59	99.13±0.04	88.05±0.08	44.62±0.37	26.22±0.25	26.12±0.31
CFL	60.58±0.15	28.55±0.30	98.70±0.01	61.40±0.51	44.19±0.69	29.62±0.43	33.47±0.68
IFCA	72.84±4.80	58.98±2.38	99.10±0.06	70.12±0.13	34.86±1.02	19.93±0.59	26.68±0.16
FedNH	50.82±0.33	26.26±0.36	98.85±0.29	56.38±0.17	32.98±0.88	17.04±0.07	24.24±0.76
FedUV	88.11±0.13	62.72±0.28	99.25±0.09	88.59±0.09	46.80±0.20	28.09±0.06	25.45±0.03
cwFedAvg (Output)	**91.23±0.04**	**67.50±0.14**	**99.52±0.03**	88.65±0.19	**56.29±0.18**	**41.38±0.12**	**43.51±0.14**

cwFedAvg | Experiments(Quantitative)

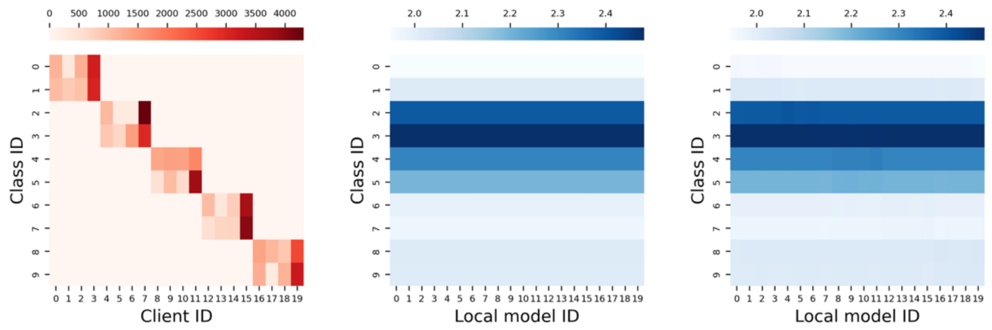
cwFedAvg outperforms under various conditions while maintaining the same communication overhead.

Table 2. Communication cost formulation and classification accuracy (%) across different settings for CIFAR-100. Σ denotes total model parameters, and C denotes the number of clusters. `cwFedAvg` (Output) denotes `cwFedAvg` selectively applied to the output layer.

Algorithm	Comm. Cost	Number of clients		Data heterogeneity		
		50 Clients	100 Clients	α=0.01	α=0.5	α=1.0
`FedAvg`	$2 \cdot \Sigma$	32.63±0.34	32.32±0.30	28.00±0.92	36.18±0.28	36.75±0.34
`FedProx`	$2 \cdot \Sigma$	33.22±0.20	32.64±0.21	27.89±0.24	35.93±0.31	36.65±0.39
`FedAMP`	$2 \cdot \Sigma$	44.97±0.27	41.37±0.35	73.46±0.40	25.41±0.14	21.23±0.40
`FedFomo`	$(1+M) \cdot \Sigma$	42.62±0.62	38.62±0.08	71.30±0.03	25.43±0.58	18.95±0.34
`CFL`	$2 \cdot \Sigma$	32.83±0.78	32.88±0.23	27.67±0.17	38.32±0.47	36.80±0.07
`IFCA`	$(1+C) \cdot \Sigma$	29.17±0.20	26.56±0.45	53.89±3.58	25.87±0.57	22.27±1.14
`FedNH`	$2 \cdot \Sigma$	33.14±0.46	32.73±0.24	25.48±0.25	37.13±0.41	20.41±0.15
`FedUV`	$2 \cdot \Sigma$	44.30±0.14	40.91±0.22	72.67±0.12	27.23±0.25	37.41±0.34
`cwFedAvg` (Output)	$2 \cdot \Sigma$	**55.90±0.35**	**53.54±0.79**	**75.20±0.21**	**40.78±0.93**	**37.50±0.10**

cwFedAvg | Experiments(Qualitative)

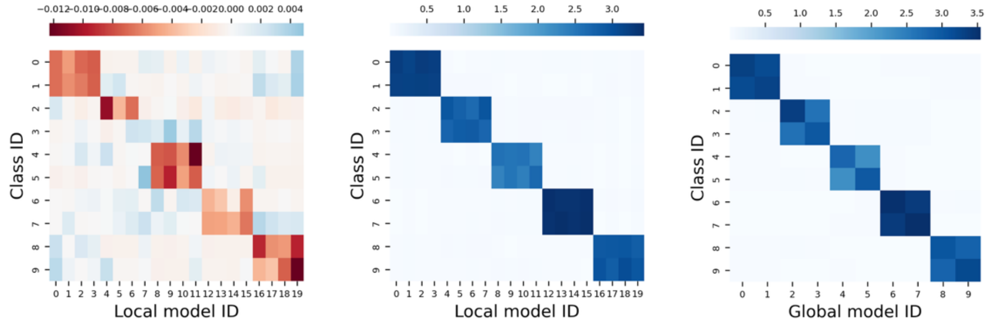
Output layer weight distribution visualization reveals how cwFedAvg achieves personalization



(a) Data distribution of clients

(b) Local models of `FedAvg`

(c) Local models of fine-tuned `FedAvg`

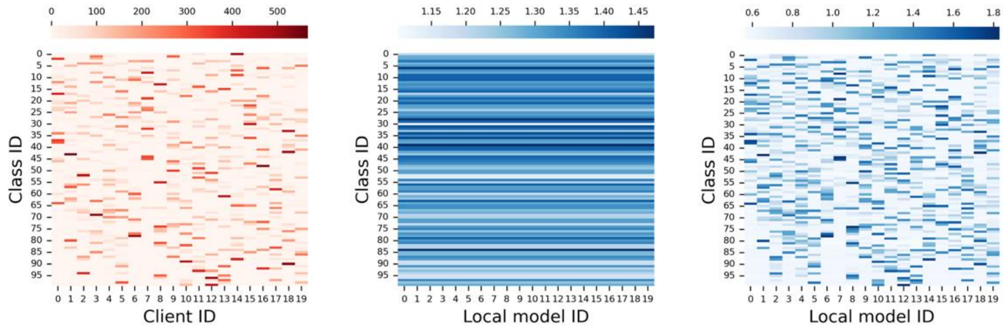


(d) Difference between (b) and (c)

(e) Local models of `cwFedAvg` w/ `WDR`

(f) Global models of w/ `WDR`

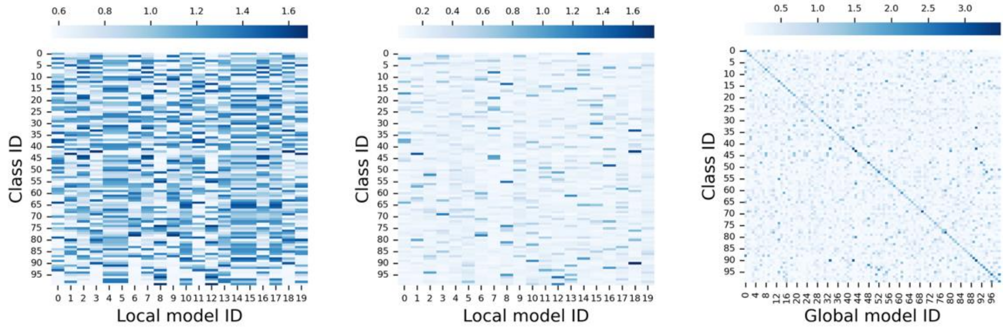
<CIFAR-10 Pathological Setting>



(a) Data distribution of clients

(b) Local models of `FedAvg`

(c) Local models of `FedAMP`



(d) Local models of `IFCA`

(e) Local models of `cwFedAvg` w/ `WDR`

(f) Global models of `cwFedAvg` w/ `WDR`

<CIFAR-100 Practical Setting(0.1)>

Thank you for watching

Class-Wise Federated Learning for
Efficient Personalization

Project Page: https://github.com/regulationLee/cwFedAvg