Connecting Low-Loss Subspace for Personalized Federated Learning

Seok-Ju Hahn[1], Minwoo Jeong[2], Junghye Lee[1]
[1]Ulsan National Institute of Science and Technology (UNIST), [2]Kakao Enterprise

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

kakaoenterprise

PAPER CODE

Personalized Federated Learning (PFL)

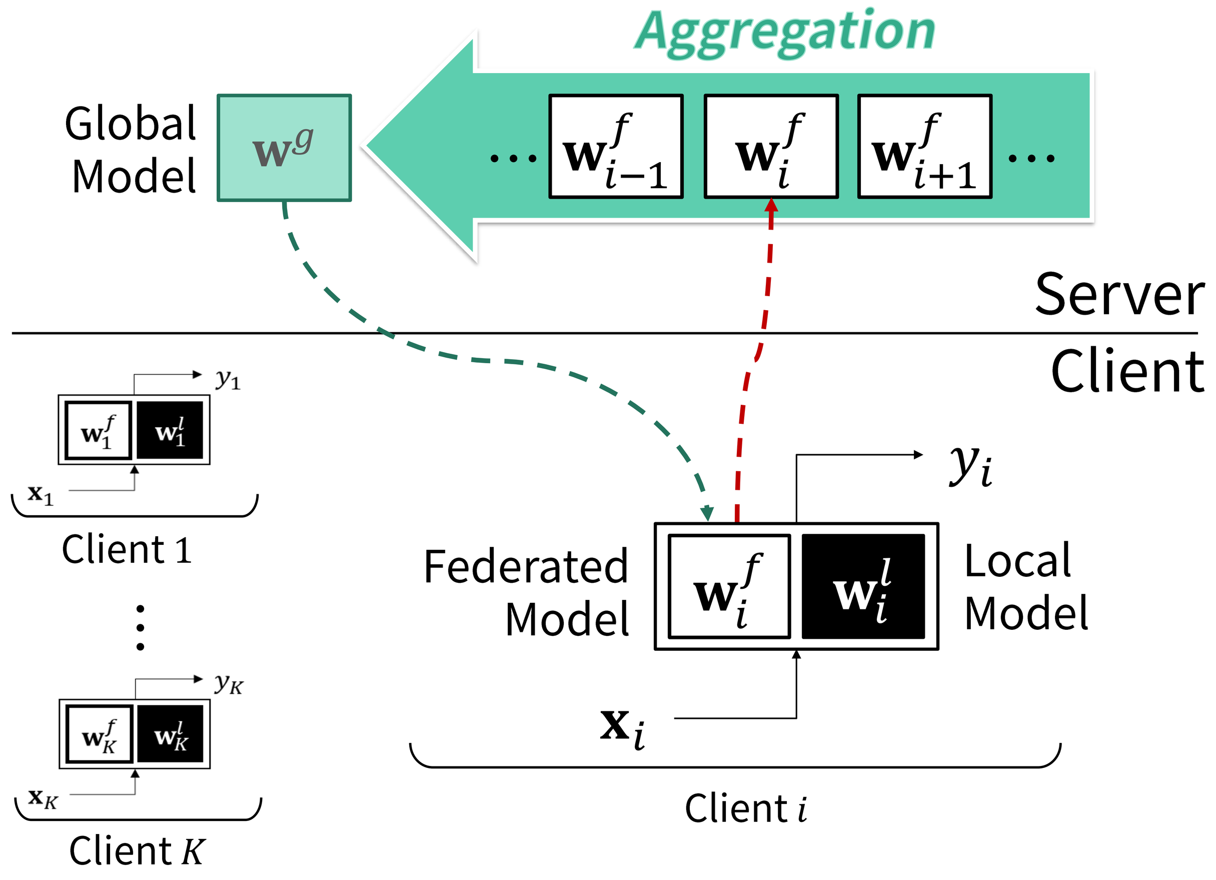
- The global objective of PFL

$$\min_{h_k \in \mathcal{H}} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\mathcal{D}_k}(h_k, \mathcal{W}_k)$$

- The local objective of PFL

$$\min_{\mathcal{W}_k} \frac{1}{n_k}\sum_{i=1}^{n_k} l\left(h_k\left(\mathbf{x}_i^{(k)}; \mathcal{W}_k\right), \mathbf{y}_i^{(k)}\right) + \Omega(\mathcal{W}_k)$$

Model Mixture-based PFL



Connectivity

- Two deep networks' local optimum CAN BE CONNECTED IN A PATH, and all solutions on the path have low-loss. (i.e., different optima can be connected to be a FLAT MINIMA – (mode) connectivity)

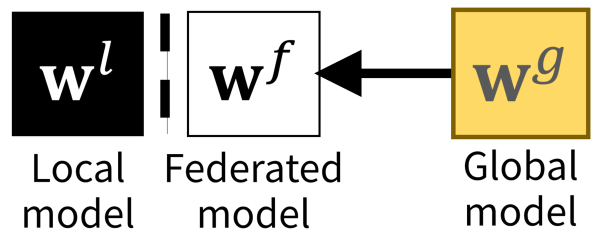
(Garipov et al., 2018; Draxler et al., 2018 ; Frankle et al., 2020)

- When the optimization trajectories of two or more deep networks are dissimilar (i.e., low cosine similarity), the ensemble performance is far higher.

(Fort et al., 2019; Wortsman et al., 2021)

SuPerFed: Proposed Method

1) Receive a global model from server as a federated model. (a local model has the same structure)

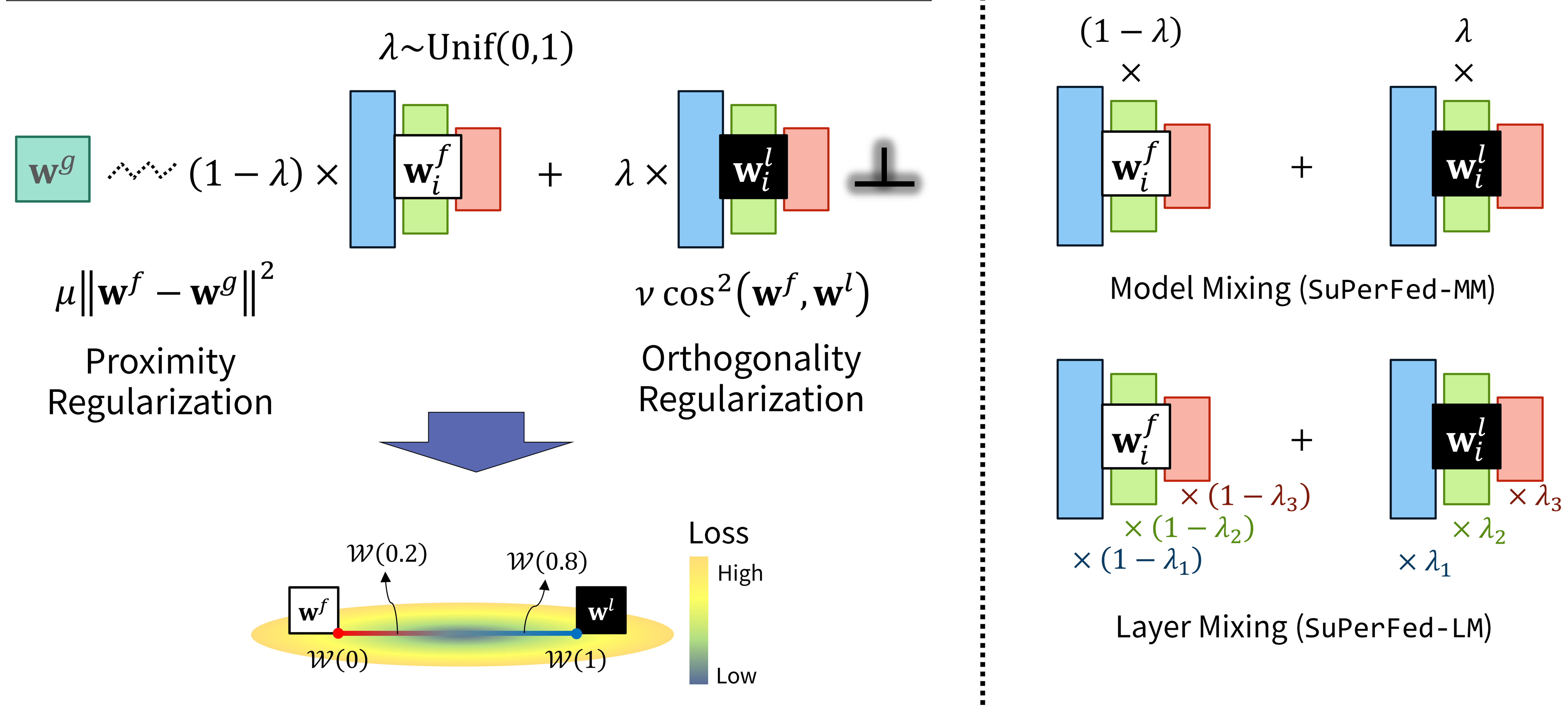


2) Mix two models instantly per every mini-batch $(\mathbf{x}, \mathbf{y})$.

$$\mathcal{W}(\lambda) = (1-\lambda)\mathbf{w}^f + \lambda\mathbf{w}^l, \lambda \sim \text{Unif}(0,1)$$

3) Update each endpoint $\mathbf{w}^f$, $\mathbf{w}^l$ at the same time with regularizations to induce the **connectivity**, while not derailing from the global update.

$$\mathcal{L}(h, \mathcal{W}) = l(h(\mathbf{x}), \mathbf{y}; \mathcal{W}(\lambda)) + \mu\left\|\mathbf{w}^f - \mathbf{w}^g\right\|^2 + \nu\cos^2(\mathbf{w}^f, \mathbf{w}^l)$$



Experimental Results

Pathological Non-IID (McMahan et al., 2017)

Dataset	MNIST			CIFAR10		
# clients	50	100	500	50	100	500
# samples	960	480	96	800	400	80
FedAvg	95.69 ± 2.39	89.78 ± 11.30	96.04 ± 4.74	43.09 ± 24.56	36.19 ± 29.54	47.90 ± 25.05
FedProx	95.13 ± 2.67	93.25 ± 6.12	96.50 ± 4.52	49.01 ± 19.87	38.56 ± 28.11	48.60 ± 25.71
SCAFFOLD	95.50 ± 2.71	90.58 ± 10.13	96.60 ± 4.26	43.81 ± 24.30	36.31 ± 29.42	40.27 ± 26.90
LG-FedAvg	98.21 ± 1.28	97.52 ± 2.11	96.05 ± 5.02	89.03 ± 4.53	70.25 ± 35.66	78.52 ± 11.22
FedPer	99.23 ± 0.66	99.14 ± 0.93	98.67 ± 2.61	89.10 ± 5.41	87.99 ± 5.70	82.35 ± 9.85
APFL	99.40 ± 0.58	99.19 ± 0.92	98.98 ± 2.22	92.83 ± 3.47	91.73 ± 4.61	87.38 ± 9.39
pFedMe	81.10 ± 8.52	82.48 ± 7.62	81.96 ± 12.28	92.97 ± 3.07	92.07 ± 5.05	88.30 ± 8.53
Ditto	97.07 ± 1.38	97.13 ± 2.06	97.20 ± 3.72	85.53 ± 6.22	83.01 ± 5.62	84.45 ± 10.67
FedRep	99.11 ± 0.63	99.04 ± 1.02	97.94 ± 3.37	82.00 ± 5.41	81.27 ± 7.90	80.66 ± 11.00
SuPerFed-MM	99.45 ± 0.46	99.38 ± 0.93	99.24 ± 2.12	94.05 ± 3.18	93.25 ± 3.80	90.81 ± 9.35
SuPerFed-LM	99.48 ± 0.54	99.31 ± 1.09	98.83 ± 3.02	93.88 ± 3.55	93.20 ± 4.19	89.63 ± 11.11

Dirichlet distribution-based non-IID (Hsu et al., 2019)

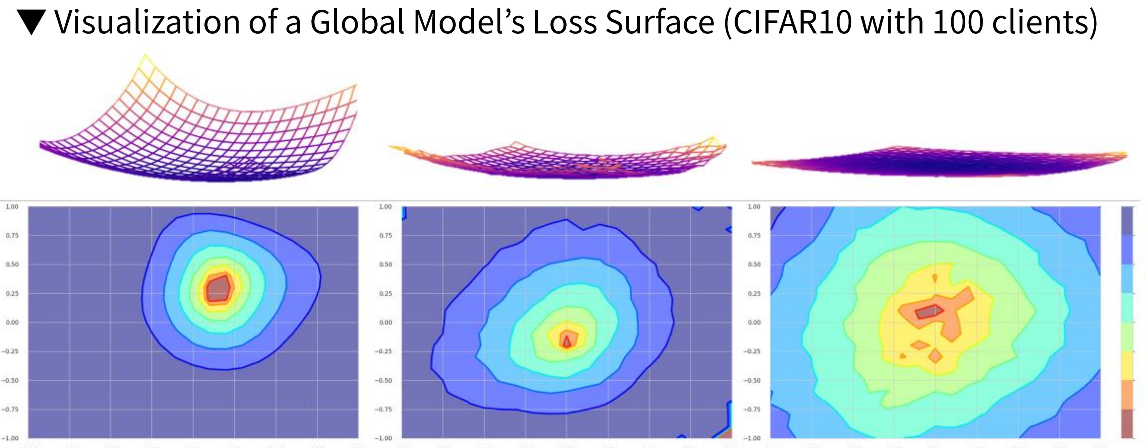
Dataset	CIFAR100			TinyImageNet		
# clients	100			200		
concentration (α)	1	10	100	1	10	100
FedAvg	58.12 ± 7.06	59.04 ± 7.19	58.49 ± 5.27	46.61 ± 5.64	48.90 ± 5.50	48.90 ± 5.40
FedProx	57.71 ± 6.79	58.24 ± 5.94	58.75 ± 5.56	47.37 ± 5.94	47.73 ± 5.94	48.97 ± 5.02
SCAFFOLD	51.16 ± 6.79	51.40 ± 5.22	52.90 ± 4.89	46.54 ± 5.49	48.77 ± 5.49	48.27 ± 5.32
LG-FedAvg	28.88 ± 5.64	21.25 ± 4.64	20.05 ± 4.61	14.70 ± 3.84	9.86 ± 3.13	9.25 ± 2.89
FedPer	46.78 ± 7.63	35.73 ± 6.80	35.52 ± 6.58	21.90 ± 4.71	11.10 ± 3.19	9.63 ± 3.12
APFL	61.13 ± 6.86	56.90 ± 7.05	55.43 ± 5.45	41.98 ± 5.94	34.74 ± 5.14	34.23 ± 5.07
pFedMe	19.00 ± 5.37	17.94 ± 4.72	18.28 ± 3.41	6.05 ± 2.84	8.01 ± 2.92	7.69 ± 2.41
Ditto	60.04 ± 6.82	58.55 ± 7.12	58.73 ± 5.39	46.36 ± 5.44	43.84 ± 5.44	43.11 ± 5.35
FedRep	38.49 ± 6.65	26.61 ± 5.20	24.50 ± 4.21	18.67 ± 4.66	9.23 ± 2.84	8.09 ± 2.83
SuPerFed-MM	60.14 ± 6.24	58.32 ± 6.25	59.08 ± 5.12	50.07 ± 5.73	49.86 ± 5.03	49.73 ± 4.84
SuPerFed-LM	62.50 ± 6.34	61.64 ± 6.23	59.05 ± 5.59	47.28 ± 5.19	48.98 ± 4.79	49.29 ± 4.82

LEAF Benchmark (Caldas et al., 2018)

Dataset	FEMNIST		Shakespeare	
# clients	730		660	
Accuracy	Top-1	Top-5	Top-1	Top-5
FedAvg	80.12 ± 12.01	98.74 ± 2.97	50.90 ± 7.85	80.15 ± 7.87
FedProx	80.23 ± 11.88	98.73 ± 2.94	51.33 ± 7.54	80.31 ± 6.95
SCAFFOLD	80.03 ± 11.78	98.85 ± 2.77	50.76 ± 8.01	80.43 ± 7.09
LG-FedAvg	50.84 ± 20.97	75.11 ± 21.49	33.88 ± 10.28	62.84 ± 13.16
FedPer	73.79 ± 14.10	86.39 ± 14.70	45.82 ± 8.10	75.68 ± 9.25
APFL	84.85 ± 8.83	98.83 ± 2.73	54.08 ± 8.31	83.32 ± 6.22
pFedMe	5.98 ± 4.55	24.64 ± 9.43	32.29 ± 6.64	63.12 ± 8.00
Ditto	64.61 ± 31.49	81.14 ± 28.56	49.04 ± 10.22	78.14 ± 12.61
FedRep	59.27 ± 15.72	70.42 ± 15.82	38.15 ± 9.54	68.65 ± 12.50
SuPerFed-MM	85.20 ± 8.40	99.16 ± 2.13	54.52 ± 7.54	84.27 ± 6.00
SuPerFed-LM	83.36 ± 9.61	98.81 ± 2.58	54.52 ± 7.54	83.97 ± 5.72

▲ Personalization

▼ Visualization of a Global Model's Loss Surface (CIFAR10 with 100 clients)



FedAvg SuPerFed-MM SuPerFed-LM

▼ Effects of Phase-wise Learning (CIFAR10 with 100 clients)

L/R	0.0	0.2	0.4	0.6	0.8
SuPerFed-MM	88.10 ± 6.25 0.91 ± 0.41	**92.13 ± 4.95** 0.92 ± 0.47	91.99 ± 4.60 **0.86 ± 0.38**	91.64 ± 6.78 1.02 ± 0.44	86.84 ± 14.86 2.79 ± 0.76
SuPerFed-LM	90.89 ± 4.58 0.80 ± 0.43	91.78 ± 4.53 0.68 ± 0.35	92.10 ± 4.41 **0.66 ± 0.32**	**92.15 ± 4.03** 1.88 ± 0.75	89.60 ± 11.56 5.23 ± 2.04

▼ Ablation Studies (CIFAR10 with 100 clients; Left: SuPerFed-MM, Right: SuPerFed-LM)