Supplementary Material for Federated Learning for Face Recognition via Intra-subject Self-supervised Learning

Hansol Kim[1,2]
hans.kim@kakaobank.com
Hoyeol Choi[1]
con.choi@kakaobank.com
Youngjun Kwak[†,1,3]
vivaan.yjkwak@kakaobank.com,yjk.kwak@kaist.ac.kr

[1] KakaoBank Corp., South Korea
[2] Graduate School of Software Technology, Kookmin University, South Korea
[3] Department of Electrical Engineering, KAIST, South Korea

Appendix

A Congervence Analysis

In this section, we analyze the convergence of FedFS. To do the analysis, we first need to make some preliminary definitions. The point where the local model is trained is designated as e. (For example, w^{e+1} means the model parameter that has completed the $e+1$th training.) We denote the loss function of FedFS as F. We do not display ψ separately, because the parameter is not updated.

Assumption 1. Lipschitz Smoothness.
If the gradient of the local model of any client c is L-Lipschitz smooth, the following formula holds.

$$||\nabla_w F(w,\theta^1) - \nabla_w F(w,\theta^2)||_2 <= L||\theta^1 - \theta^2||_2 \tag{A.1}$$

$$||\nabla_\theta F(w^1,\theta) - \nabla_\theta F(w^2,\theta)||_2 <= L||w^1 - w^2||_2 \tag{A.2}$$

Assumption 2. Unbiased Gradient and Bounded Variance
The w and θ parameters each use SGD as an optimization function, so they each have unbiased and bounded variance. The parameter update process using SGD is as follows:

$$w^{e+1} = w^e - \eta\nabla_w F(w^e,\theta^e) \tag{A.3}$$

$$\theta^{e+1} = \theta^e - \eta\nabla_\theta F(w^e,\theta^e) \tag{A.4}$$

† Corresponding author

Assuming that the amount of change in the parameter is within a certain range, the conditions are as follows:

$$||w^e - w^*||_2 <= \varepsilon_w \tag{A.5}$$
$$||\theta^e - \theta^*||_2 <= \varepsilon_\theta \tag{A.6}$$

where ε_w and ε_θ are positive value. With this assumption, updates to each parameter occur randomly within a certain range.

Theorem 1. Convergence analysis
Based on Assumption 1 and Assumption 2, a convergence analysis of FedFS can be performed as follows:

$$||w^{e+1} - w^*||_2 <= \varepsilon_w - \eta L \varepsilon_\theta \tag{A.7}$$
$$||\theta^{e+1} - \theta^*||_2 <= \varepsilon_\theta - \eta L \varepsilon_w \tag{A.8}$$

Based on the above analysis, we confirm that the parameters of FedFS converge within the real number range.

Proof 1.
First, the change in the objective function is estimated using the distance between parameters as follows.

$$||w^{e+1} - w^*||_2 = ||w^e - \eta \nabla_w F(w^e, \theta^e) - w^*||_2 \tag{A.9}$$
$$= ||w^e - w^* - \eta \nabla_w F(w^e, \theta^e)||_2 \tag{A.10}$$
$$= ||w^e - w^* - \eta (\nabla_w F(w^e, \theta^e) - \nabla_w F(w^*, \theta^*))||_2 \tag{A.11}$$
$$<= ||w^e - w^*||_2 - \eta ||(\nabla_w F(w^e, \theta^e) - \nabla_w F(w^*, \theta^*)||_2 \tag{A.12}$$
$$<= ||w^e - w^*||_2 - \eta L ||\theta^e - \theta^*||_2 \tag{A.13}$$

Similar to the expansion process above, the same process is repeated for w to obtain the follows:

$$||\theta^{e+1} - \theta^*||_2 <= ||\theta^e - \theta^*||_2 - \eta L ||w^e - w^*||_2 \tag{A.14}$$

Now, based on Assumption 2, we develop the following:

$$||w^{e+1} - w^*||_2 <= ||w^e - w^*||_2 - \eta L ||\theta^e - \theta^*||_2 \tag{A.15}$$
$$<= \varepsilon_w - \eta L \varepsilon_\theta \tag{A.16}$$
$$||\theta^{e+1} - \theta^*||_2 <= ||\theta^e - \theta^*||_2 - \eta L ||w^e - w^*||_2 \tag{A.17}$$
$$<= \varepsilon_\theta - \eta L \varepsilon_w \tag{A.18}$$

Through this, Theorem 1 was proven.

B Ablation Studies

Setup	Modules		DigiFace-1M	VGGFace
	Regularize loss	Adaptive soft label	TPIR@FPIR=0.001	TPIR@FPIR=0.001
Centrally trained with PocketNet			0.9128	0.9806
A	×	×	0.9051	0.9645
B	✓	×	0.9537	0.9902
Ours(C)	✓	✓	0.9794	0.9934

Table B.1: The TPIR performance for ablation studies

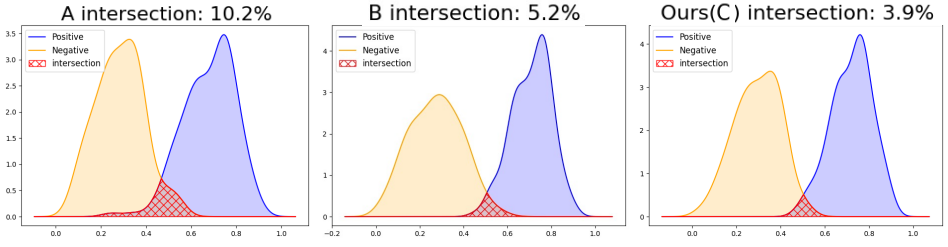

Figure B.1: Average similarity distribution of clients participated in federated learning

Intra-subject self-supervised learning learning defined in the Equation 9, improves performance of personalized face recognition compared to previous approaches. We go further and check whether the proposed method reduces intra-class variation and analyze how the method affects performance. We set the participation rate at 0.7, and use PocketNet [5] as the pre-trained model. As shown in Figure B.1 and Table B.1, we can see that the intra-subject self-supervised learning method considering correlation shows superior performance compared to using the general entropy learning method and regularize loss has a significant impact on performance by preventing overfitting and bias. In particular, as shown in the figure B.1, the proposed method has the smallest intersection of the positive similarity area and the negative similarity area. Through these results, we confirm that intra-subject self-supervised learning reduces intra-class variance.