

SUPPLEMENTARY MATERIALS FOR “PERSONALIZED SEMANTICS EXCITATION”

Anonymous authors

Paper under double-blind review

1 RELATED WORKS

1.1 FEDERATED LEARNING

This scenario aims to learn a generic model for all distributed nodes via their collaboration when avoiding private data leakage. The classical and representative FedAvg McMahan et al. (2017) proposes to averagely integrate the client models in the server as the global model and conduct multiple communication between server and clients to gradually learn high-generalization model. The sequel advances FedAvg in either server model integration Chen & Chao (2020); Lin et al. (2020); Yurochkin et al. (2019) or local training strategy Malinovskiy et al. (2020); Yuan & Ma (2020); Durmus et al. (2021). Concretely, to attain the synchronous optimization across various clients, FedProx Li et al. (2018) designs the specific regularization into the global model. Similarly, SCAFFOLD Karimireddy et al. (2019) attempts to adjust the network parameters to control local gradients. These FL methods indeed achieve knowledge sharing across different clients and assist certain clients with insufficient training samples to obtain better performance. However, the generic global model fails to produce identical positive influence on all clients when there exists considerable data distribution divergence across various clients. This practical challenge motives the exploration of personalized federated learning (PFL).

1.2 PERSONALIZED FEDERATED LEARNING

To overcome the FL problem, the efficiency adaptation manner is to fine-tune the generic model received from the server with the local private data Yu et al. (2020); Arivazhagan et al. (2019). On one hand, Ditto Li et al. (2021) develops the ℓ_2 -norm regularization over the difference between global and local model parameters to customize the local network while indiscriminately preserving extensive global knowledge. Similarly, T Dinh et al. (2020); Hanzely et al. (2020) also introduce other regularizers to mitigate the model shift during local training stage. On the other hand, the mixture of global and local networks has been a promising personalized manner Zec et al. (2020); Deng et al. (2020). Due to the recent success of meta-learning on model generalization, several works share the same spirit to learn the meta-model across different clients and then gradually adapt it to each specific client Fallah et al. (2020). Moreover, FedRep Collins et al. (2021) conducts more explicit personalization by decomposing the network architecture into global feature extractor and private classifier. In fact, these methods perform model customization from the identical generic model for all clients. Another direction is to first cluster clients into multiple groups and train group-wise global models in the server Huang et al. (2021); Zhang et al. (2020).

Although the mentioned PFL strategies achieve appealing performances, they cannot clearly illustrate why their model can conduct better model customization. To answer this question, we first attempt to visualize the discriminative information learned by them over the original images and empirically notice that certain methods are likely to discard important semantic when adapting from global model. Thus, we claim that the ideal customization not only discovers the well-learned discriminative knowledge from global model but also preserves personal information from local data to promote the robustness of local model. With this motivation, this work proposes active knowledge imitation and enhancement to better solve PFL challenges.

2 THEORETICAL ANALYSIS

The objective function of our PSE:

$$\min_{\tilde{\Theta}_g, \tilde{\Theta}_c} \tilde{\mathcal{L}} = \underbrace{\sum_j \mathcal{L}(\tilde{\mathbf{p}}_j, y_j) + \mathcal{L}(\mathbf{p}_j, y_j)}_{\text{Obj 1}} + \underbrace{\sum_l \sum_k \frac{\xi}{2} \cdot \mathbb{I}(\Delta^{k_l} \geq \bar{\Delta}) \cdot \|\tilde{\mathbf{W}}_k^{(l)} - \mathbf{W}_k^{(l)}\|_{\ell_2}^2}_{\text{Obj 2}}. \quad (1)$$

Assumption 1. In each client, the stochastic gradient $\mathbf{g}_t = \nabla \tilde{\mathcal{L}}(\tilde{\Theta}_t, \mathbf{x}_t)$ at time t is an unbiased estimator of the local gradient with the expectation as $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{g}_t] = \nabla \tilde{\mathcal{L}}(\tilde{\Theta}_t) = \nabla \tilde{\mathcal{L}}_t$ and variance as $\mathbb{E}[\|\mathbf{g}_t - \nabla \tilde{\mathcal{L}}_t\|_2^2] \leq \delta^2$.

Assumption 2. The objective function optimized in each client is L_1 -Lipschitz smooth. In other words, the gradient of Eq. (1) is L_1 -Lipschitz continuous Malherbe & Vayatis (2017), i.e., $\|\nabla \tilde{\mathcal{L}}_{t_1} - \nabla \tilde{\mathcal{L}}_{t_2}\|_2 \leq L_1 \|\tilde{\Theta}_{t_1} - \tilde{\Theta}_{t_2}\|_2$, where $\mathcal{L}_{t_{1/2}}$ means the loss values at local iteration time $t_{1/2}$.

Theorem 1. When assumption 1 and 2 hold, we have the following conclusion in any arbitrary client after per communication round (r):

$$\mathbb{E}[\tilde{\mathcal{L}}_{(r+1)\tau}] \leq \tilde{\mathcal{L}}_{r\tau+1} - \left(\eta - \frac{L_1\eta^2}{2}\right) \sum_{e=1}^{\tau-1} \|\nabla \tilde{\mathcal{L}}_{r\tau+e}\|_2^2 + \frac{L_1\tau\eta^2}{2} \delta^2, \quad (2)$$

where τ is the total iteration of local model update and η is the learning rate. This theorem suggests that selecting appropriate η can achieve our expected gradient decrease in one communication round so that it finally can guarantee the convergence of model.

Proof. Given the assumption 2, we can rewrite the L_1 -Lipschitz continuous condition as:

$$\tilde{\mathcal{L}}_{t_1} - \tilde{\mathcal{L}}_{t_2} \leq \langle \nabla \tilde{\mathcal{L}}_{t_2}, (\tilde{\Theta}_{t_1} - \tilde{\Theta}_{t_2}) \rangle + \frac{L_1}{2} \|\tilde{\Theta}_{t_1} - \tilde{\Theta}_{t_2}\|_2^2. \quad (3)$$

With this formulation, we can deduce the upper bounder of loss function after one iteration by allowing $\tilde{\Theta}_{r\tau+2} = \tilde{\Theta}_{r\tau+1} - \eta \mathbf{g}_{r\tau+1}$ as the following:

$$\begin{aligned} \tilde{\mathcal{L}}_{r\tau+2} &\leq \tilde{\mathcal{L}}_{r\tau+1} + \langle \nabla \tilde{\mathcal{L}}_{r\tau+1}, (\tilde{\Theta}_{r\tau+2} - \tilde{\Theta}_{r\tau+1}) \rangle + \frac{L_1}{2} \|\tilde{\Theta}_{r\tau+2} - \tilde{\Theta}_{r\tau+1}\|_2^2 \\ &= \tilde{\mathcal{L}}_{r\tau+1} - \eta \langle \nabla \tilde{\mathcal{L}}_{r\tau+1}, \mathbf{g}_{r\tau+1} \rangle + \frac{L_1}{2} \|\eta \mathbf{g}_{r\tau+1}\|_2^2. \end{aligned} \quad (4)$$

And then, we calculate the expectation over both sides for the random variable \mathbf{x} with the following:

$$\mathbb{E}[\tilde{\mathcal{L}}_{r\tau+2}] \leq \tilde{\mathcal{L}}_{r\tau+1} - \eta \mathbb{E}[\langle \nabla \tilde{\mathcal{L}}_{r\tau+1}, \mathbf{g}_{r\tau+1} \rangle] + \frac{L_1\eta^2}{2} \mathbb{E}[\|\mathbf{g}_{r\tau+1}\|_2^2] \quad (5)$$

$$\leq \tilde{\mathcal{L}}_{r\tau+1} - \eta \|\nabla \tilde{\mathcal{L}}_{r\tau+1}\|_2^2 + \frac{L_1\eta^2}{2} (\|\nabla \tilde{\mathcal{L}}_{r\tau+1}\|_2^2 + \text{Var}(\mathbf{g}_{r\tau+1})) \quad (6)$$

$$\leq \tilde{\mathcal{L}}_{r\tau+1} - \left(\eta - \frac{L_1\eta^2}{2}\right) \|\nabla \tilde{\mathcal{L}}_{r\tau+1}\|_2^2 + \frac{L_1\eta^2}{2} \delta^2, \quad (7)$$

where $\text{Var}(\cdot)$ means the variance of the variable. Finally, we repeat the above inequality for τ times and achieve the conclusion.

Theorem 2. Given any ϵ , after R round communication, we infer that

$$\frac{1}{R\tau} \sum_{r=1}^{R-1} \sum_{e=1}^{\tau-1} \mathbb{E}[\|\nabla \tilde{\mathcal{L}}_{r\tau+e}\|_2^2] \leq \epsilon, \quad R \geq \frac{2(\tilde{\mathcal{L}}_1 - \tilde{\mathcal{L}}_*)}{\tau\epsilon(2\eta - L_1\eta^2) - \tau\eta^2 L_1 \delta^2}, \quad (8)$$

where $\eta < \frac{2\epsilon}{L_1(\epsilon + \delta^2)}$ and $\tilde{\mathcal{L}}_*$ denotes the loss of the optimal solution for the local model. This theorem illustrates the convergence rate of model, which is related to the overall communication round and the expectation of ℓ_2 -norm of gradient. Sufficient communication rounds make the bound tighter. Please refer to the supplementary material for the proofs of two theorems.

Table 1: Average Recognition Accuracy (%) with standard deviation under novel joint label and data shift scenarios.

Datasets	FEMNIST			FashionMNIST			FashionMNIST		
Modality	(Gray, Color)			(Gray, Color)			(Color, Edge)		
(#M, #C)	(200,3)	(200,4)	(200,5)	(100,3)	(100,4)	(100,5)	(200,3)	(200,4)	(200,5)
Local	81.97±4.37	80.40±4.20	79.44±4.49	81.46±3.55	79.62±3.14	76.95±3.15	83.53±3.26	81.95±3.84	80.57±3.25
FedAvg+FT	83.17±3.04	81.97±2.75	81.53±3.12	84.28±3.31	82.36±2.85	79.80±3.12	86.39±3.33	84.08±3.42	82.99±3.25
FedProx+FT	82.87±3.65	81.36±3.44	81.03±3.25	84.51±2.74	82.31±2.70	79.50±2.97	87.05±3.09	84.25±2.79	83.09±2.88
SCAFFOLD+FT	84.00±3.62	81.54±2.97	82.04±3.64	84.79±3.43	82.12±2.75	79.94±2.85	85.54±3.05	83.19±3.26	83.05±3.07
Fed-MTL	81.14±2.62	80.30±2.90	79.24±2.90	78.70±1.88	77.14±2.58	78.39±2.37	81.12±2.14	79.49±2.11	79.89±2.40
LG-Fed	83.27±3.18	81.40±2.87	80.03±2.87	81.59±1.85	79.23±2.73	75.89±2.27	83.86±1.96	80.90±2.33	78.31±2.06
L2GD	81.88±3.02	80.53±3.18	79.68±2.93	80.16±2.26	78.90±2.78	77.46±1.96	81.75±2.66	80.86±2.44	79.52±2.18
APFL	82.85±2.50	81.17±2.74	81.14±2.38	85.25±2.33	81.16±1.93	78.73±1.81	85.96±2.59	82.46±1.97	79.22±2.45
Ditto	85.23±2.13	82.94±2.17	82.34±2.56	88.11±2.19	85.76±1.95	84.46±1.73	87.82±1.65	84.77±2.04	84.13±1.58
FedRep	84.43±3.00	83.54±2.03	83.51±2.88	86.71±2.25	83.01±1.58	83.49±2.41	84.78±2.33	85.10±2.50	84.46±1.94
Ours	88.81±1.47	87.86±0.99	87.98±1.17	89.58±0.43	88.12±1.45	86.61±0.31	89.97±1.23	87.95±1.28	85.69±1.34

Proof. According to the conclusion in Theorem 1, we can consider conducting R round communication between server and clients and easily obtain the expectation of $\tilde{\Theta}$ on both side of Eq. 2 with the formulation as:

$$\sum_{r=1}^{R-1} \mathbb{E}[\tilde{\mathcal{L}}_{(r+1)\tau+1}] \leq \sum_{r=1}^{R-1} \tilde{\mathcal{L}}_{r\tau+1} - (\eta - \frac{L_1\eta^2}{2}) \sum_{r=1}^{R-1} \sum_{e=1}^{\tau-1} \mathbb{E}[\|\nabla \tilde{\mathcal{L}}_{r\tau+e}\|_2^2] + \frac{R\tau L_1\eta^2}{2} \delta^2. \quad (9)$$

To this end, we can obtain the following formulation:

$$\frac{1}{R\tau} \sum_{r=1}^{R-1} \sum_{e=1}^{\tau-1} \mathbb{E}[\|\nabla \tilde{\mathcal{L}}_{r\tau+e}\|_2^2] \leq \frac{\frac{1}{R\tau} \sum_{r=1}^{R-1} (\tilde{\mathcal{L}}_{r\tau+1} - \mathbb{E}[\tilde{\mathcal{L}}_{(r+1)\tau}]) + \frac{L_1\eta^2}{2} \delta^2}{\eta - L_1\eta^2}. \quad (10)$$

On the other hand, we have $\sum_{r=1}^{R-1} (\tilde{\mathcal{L}}_{r\tau+1} - \mathbb{E}[\tilde{\mathcal{L}}_{(r+1)\tau}]) < \tilde{\mathcal{L}}_1 - \tilde{\mathcal{L}}_*$. Thus, for arbitrary ϵ , we have

$$\frac{\frac{1}{R\tau} (\tilde{\mathcal{L}}_1 - \tilde{\mathcal{L}}_*) + \frac{L_1\eta^2}{2} \delta^2}{\eta - L_1\eta^2} < \epsilon. \quad (11)$$

Thus, we can achieve the the condition of communication round as

$$R \geq \frac{2(\tilde{\mathcal{L}}_1 - \tilde{\mathcal{L}}_*)}{\tau\epsilon(2\eta - L_1\eta^2) - \tau\eta^2 L_1\delta^2}. \quad (12)$$

Since the $\tau\epsilon(2\eta - L_1\eta^2) - \tau\eta^2 L_1\delta^2 > 0$, we have $\eta < \frac{2\epsilon}{L_1(\epsilon + \delta^2)}$. Theorem 2 provides the delicate convergence rate for our PSE method. Given any bound ϵ , we can select the appropriate communication round and the learning rate to adjust the rate and achieve the training convergence.

3 ADDITIONAL EMPIRICAL ANALYSIS

Due to the limited space of main manuscript, we only visualize the attention maps of a few samples, which are drawn by multiple personalized federated learning methods. To fully understand how each method conducts the customization, we show more examples in Figure 1. Compared with others, our method can capture more important discriminative information and utilize them to do the final decision. In other words, our PSE enables the model to adapt the distribution property of each client and conduct better personalization. Moreover, we also report the standard deviation of experimental results in Table 1 and Table 2 corresponding to that of manuscript.

4 DISCUSSION

Our method aims to achieve knowledge sharing across numerous distributed clients without private data leakage and customizes local model to adapt their own data distribution. Thus, our method has



Figure 1: Attention Maps drawn by multiple personalized federated learning methods.

Table 2: Recognition Accuracy (%) with standard deviation under conventional label shift scenarios.

Datasets	CIFAR-10			CIFAR-100			FashionMNIST		
(#M, #C)	(20,2)	(20,3)	(20,4)	(50,5)	(50,10)	(50,15)	(100,3)	(100,4)	(100,5)
Local	79.65±4.03	73.97±3.97	67.54±4.03	73.35±4.24	58.76±4.11	49.79±4.72	89.65±3.08	86.37±3.40	85.75±3.26
FedAvg+FT	82.94±2.43	78.23±2.98	74.62±2.26	77.01±2.26	61.96±2.22	55.40±2.32	91.43±3.30	89.00±2.68	87.36±2.64
FedProx+FT	82.44±2.18	76.74±2.44	73.63±2.41	74.10±2.60	60.40±2.12	53.35±2.42	88.35±2.93	87.05±2.76	85.51±3.37
SCAFFOLD+FT	82.03±2.90	76.51±2.11	72.92±2.59	75.09±2.71	59.92±2.30	51.54±2.51	90.33±3.41	87.68±2.65	85.22±3.08
Fed-MTL	83.19±1.59	75.81±1.53	69.57±1.99	65.28±1.96	54.84±2.02	48.72±2.12	84.65±2.35	82.59±2.42	82.86±2.20
LG-Fed	84.24±1.76	77.1±2.43	71.23±2.08	67.17±2.46	54.31±1.73	50.63±2.18	87.07±1.94	84.51±2.15	81.19±1.88
L2GD	83.76±1.62	76.26±2.09	69.8±1.73	67.15±1.88	55.30±2.08	50.12±1.73	85.50±1.99	83.88±2.23	82.84±2.28
APFL	82.09±2.30	78.80±2.23	74.29±1.74	72.81±2.05	61.77±1.99	54.04±1.90	90.58±2.65	86.83±2.31	85.67±2.04
Ditto	84.74±1.68	80.34±1.25	76.25±2.39	75.23±1.33	65.40±1.64	56.14±1.38	91.21±1.62	89.91±1.74	88.81±1.55
FedRep	84.12±2.48	80.39±2.35	76.28±2.24	78.30±1.54	63.52±2.08	58.94±2.14	92.71±1.68	90.73±1.92	89.56±2.40
Ours	86.95±0.73	82.98±1.38	78.03±1.19	79.58±1.53	67.10±1.46	62.46±1.68	94.03±1.43	91.77±0.89	90.47±0.89

positive influence on solving data privacy protection and fairness, which are the raised concerns in real-world applications. Definitely, since our method conducts model transmission between clients and server, it can leak out certain information of client in implicit manner. This is a common challenge in current PFL setting, which is an important research direction. Our all experiments are performed with Pytorch platform in one Nvidia GeForce RTX 3090 (24GB).

REFERENCES

- Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.
- Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2020.
- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

- Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.
- Alp Emre Durmus, Zhao Yue, Matas Ramon, Mattina Matthew, Whatmough Paul, and Saligrama Venkatesh. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.
- Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.
- Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33:2304–2315, 2020.
- Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021.
- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. 2019.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.
- Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.
- Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.
- Cédric Malherbe and Nicolas Vayatis. Global optimization of lipschitz functions. In *International Conference on Machine Learning*, pp. 2314–2323. PMLR, 2017.
- Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pp. 6692–6701. PMLR, 2020.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.
- Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.
- Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.
- Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.
- Edvin Listo Zec, Olof Mogren, John Martinsson, Leon René Sütthof, and Daniel Gillblad. Specialized federated learning using a mixture of experts. *arXiv preprint arXiv:2010.02056*, 2020.
- Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.