ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447362

PERSONALIZED FEDERATED LEARNING WITH ATTENTION-BASED CLIENT SELECTION

Zihan Chen, Jundong Li, Cong Shen

Department of ECE, University of Virginia, Charlottesville, VA, USA

ABSTRACT

Personalized Federated Learning (PFL) relies on collective data knowledge to build customized models. However, non-IID data between clients poses significant challenges, as collaborating with clients who have diverse data distributions can harm local model performance, especially with limited training data. To address this issue, we propose FEDACS, a new PFL algorithm with an Attention-based Client Selection mechanism. FEDACS integrates an attention mechanism to enhance collaboration among clients with similar data distributions and mitigate the data scarcity issue. It prioritizes and allocates resources based on data similarity. We further establish the theoretical convergence behavior of FEDACS. Experiments on CIFAR10 and FMNIST validate FEDACS's superiority, showcasing its potential to advance personalized federated learning. By tackling non-IID data challenges and data scarcity, FEDACS offers promising advances in personalized federated learning.

***Index Terms*—** Federated Learning; Personalization; Deep Learning

1. INTRODUCTION

Federated learning is a collaborative learning paradigm that allows multiple clients to work together while ensuring the preservation of their privacy [1]. By leveraging the collective knowledge and data from all participating clients, federated learning aims to achieve better learning performance compared with individual client efforts [2]. This collaborative nature has made federated learning increasingly popular, finding numerous practical applications where data decentralization and privacy are paramount. The privacy-preserving solutions offered by federated learning have found extensive applications in domains such as healthcare, smart cities, and finance [3, 4, 5].

However, the effectiveness of the collaborative approach in federated learning is highly dependent on the data distribution among the clients. While federated learning performs exceptionally well when data distribution among clients is independent and identically distributed (IID), this is not the case in many real-world scenarios [6]. When the global model, which is collectively trained across decentralized clients, encounters diverse datasets with varying statistical characteristics, it may face challenges in effectively generalizing to the unique local data of each client [7, 8]. The performance of the global model may be suboptimal on specific clients' data due to differences in data distributions and patterns. This limitation becomes more pronounced as the diversity among local data from different clients continues to increase [9].

To address this issue, personalized federated learning (PFL) techniques have been proposed, which aim to overcome the limitations of traditional federated learning by incorporating personalized adaptation in local clients. Instead of relying solely on a single global model, PFL allows each client to tailor the model to her specific data distribution, preferences, and local context [10]. Most existing PFL methods focus on taking the global model as an initial model for local training or incorporating the global model into the loss function [9, 11, 12, 13, 14]. However, recent work by Huang et al. [15] challenges the assumption that a single global model can adequately fit all clients in personalized cross-silo federated learning with non-IID data. In particular, they argue that the misassumption of a universally applicable global model remains a fundamental bottleneck [15]. However, their claims lack substantial theoretical foundations or sufficient empirical evidence. To shed light on this issue, Mansour et al. [16] provide theoretical insights into the generalization of the uniform global model and demonstrate that the discrepancy between local and global distributions influences the disparity between local and global models. This study presents an empirical example to further validate that a single global model cannot adequately fit all clients. Table 1 presents the learning performance of local training models and two PFL methods [13, 11]. For each dataset, we utilize a Dirichlet distribution with the parameter of 0.5 to partition it, and subsequently, we sample 50 data points from the training dataset for the training process. Specifically, DITTO-P and PERAVG-P represent the results of personalized models, while DITTO-G and PERAVG-G reflect the performance of global models generated by the respective algorithms. It is evident from the table that the global models produced by the PFL methods struggle to achieve satisfactory generalization across clients and, in some cases, even perform worse than the locally trained models. This empirical example highlights the need for novel approaches beyond the simple fine-tuning of global models and accounts for the inherent heterogeneity in data distributions among clients.

Table 1: Test accuracy of global and personalized models.

Dataset	CIFAR10 dir = 0.5	FMNIST dir = 0.5
Local	77.29	75.98
DITTO-P	77.33	78.79
DITTO-G	39.47	76.34
PERAVG-P	73.07	74.07
PERAVG-G	28.90	69.14

Motivated by the insights gained from [15] and recognizing the challenges posed by the scarcity of data in specific areas such as medical care, in this paper, we propose a PFL algorithm with an attention-based client selection mechanism (FEDACS) that aims to overcome the dependency on a single global model and leverage the model information from other clients to develop a personalized model

for each client. By employing an attention mechanism, FEDACS allows clients to receive and process personalized messages tailored to their specific model parameters, enhancing the effectiveness of collaboration. Additionally, we provide the convergence of our proposed FEDACS method in general scenarios, ensuring the reliability and stability of the algorithm. This demonstrates that the proposed method can effectively converge to optimal solutions for personalized models in general settings. To evaluate the effectiveness of the proposed methods, we conduct extensive experiments on various datasets and settings, allowing for a comprehensive performance evaluation. The results of the experiments demonstrate the superior performance of the proposed methods compared to those of existing approaches.

2. PERSONALIZED FEDERATED LEARNING

2.1. Formulation

Standard federated learning (FL) aims to train a global model by aggregating local models contributed by multiple clients without the need for their raw data to be centralized. In FL, there are n clients communicating with a server to solve the problem: $\min_w \frac{1}{n}\sum_{i=1}^{n}\mathbb{F}_i(w)$ to find a global model w. The function $\mathbb{F}_i$ denotes the expected loss over the data distribution of the client i. One of the first FL algorithms is FEDAVG [2], which uses local stochastic gradient descent (SGD) updates and builds a global model from a subset of clients.

However, the performance of the global model tends to degrade when faced with clients whose data distributions differ significantly from the global training data distribution. On the other hand, local models trained on the respective data distributions match the distributions at inference time. Still, their generalization capabilities are limited due to the data scarcity issue [16]. Personalized federated learning (PFL) can be viewed as a middle ground between pure local models and global models, as it combines the generalization properties of the global model with the distribution matching characteristic of the local model [11, 13, 12, 9, 14].

In PFL, the goal is to train personalized models $w_1, w_2, ..., w_n$ for individual clients while respecting data privacy and accommodating user-specific requirements. The problem can be framed as:

$$w_1^*, w_2^*, ..., w_n^* = \underset{w_1,...,w_n}{\operatorname{argmin}} \sum_{i=1}^{n}\mathbb{F}_i(w_i).$$

Instead of solving the traditional PFL problem, Huang et al. [15] argued that the fundamental bottleneck in personalized cross-silo federated learning with non-IID data is that the misassumption of one global model can fit all clients. Therefore, they proposed a different approach by adding a regularized term with ℓ_2-norm of model weights distances among clients and formulate the PFL problem as:

$$\min_W \mathcal{F}_\lambda(W) = \mathcal{F}(W)+\lambda\mathcal{R}(W) = \mathcal{F}(W)+\lambda\sum_{i,j=1}^{n}\mathcal{R}(||w_i-w_j||^2), \tag{1}$$

in which $W = [w_1, ..., w_n]$ is a matrix that contains clients' local models as its columns, and λ is a regularization parameter. $\mathcal{F}(W) = \sum_{i=1}^{n}\mathbb{F}_i(w_i)$ is known as "client-side personalization" in the context of PFL. $\mathcal{R}$ is a regularization term designed to help clients with similar data distributions collaboratively train their own models.

2.2. Our proposed method: FEDACS

We propose a novel federated attention-based client selection algorithm (FEDACS) (as illustrated in Algorithm1) to solve the optimization problem (1). In this paper, we mainly consider:

$$\mathcal{R}(W) = \sum_{i,j=1}^{n} s_{ij}||w_i - w_j||^2, \tag{2}$$

where s_{ij} is the normalized score to measure data distribution similarity. It is worth mentioning that the formation of $\mathcal{R}(W)$ is easy to generate into other functions, like $\mathcal{R}(x) = 1 - e^{-x^2/\sigma}$ [15]. The above formulation implies that locally optimizing the objective function requires each client to collect the other clients' model parameters for computation. However, researchers find that adversaries can infer data information in the training set based on the model parameters, and directly gathering model parameters will violate the principles of federated learning to protect data privacy [17, 18, 19]. To address this dilemma, we leverage the idea of incremental optimization [20]. Specifically, we iteratively optimize $\mathcal{F}_\lambda(W)$ by alternatively optimizing $\mathcal{R}(W)$ and $\mathcal{F}(W)$ until convergence. Note that FEDACS handles the refinement of $\mathcal{R}(W)$ on the server side. We introduce an intermediate model u_i for each client, and these intermediate models form the model matrix $U = [u_1, .., u_n]$ in the same way as W. In the k-th iteration, we first apply a gradient descent step to update U:

$$U^k = W^{k-1} - \alpha_k\nabla\mathcal{R}(W^{k-1}), \tag{3}$$

where W^{k-1} is the collection of local models from the last round and α_k is the learning rate. Under the definition (2), the gradient of $\mathcal{R}(W)$'s i-th column is $\nabla\mathcal{R}(W)_i = w_i - \sum_j s_{ij}w_j$. In this paper, we use the cosine similarity of the model parameters instead of the data similarity to avoid the accessibility of the server to local data, i.e. $s_{ij} = \frac{\langle w_i, w_j\rangle}{||w_i||||w_j||}$.

As mentioned above, the performance of a local model could be adversely affected if each client averages information from those with different data distributions. Thus, in practice, we also introduce δ to control the degree of collaboration. We adopt a dynamic strategy to define δ for each round. After gathering the parameters from clients at the k-th round, the server first computes the similarity of the model s_{ij}^k to generate the similarity matrix $\mathcal{S}^k : (\mathcal{S}^k)_{ij} = s_{ij}^k$, then δ is calculated by the p-quantile of all elements in $\mathcal{S}^k$. We can adjust the ratio p to alter the degree of collaboration. Modifying δ can effectively mitigate the aggregation effect from initial training rounds, where the model parameters fail to represent the underlying data distribution accurately. After having δ^k, we only consider the model similarity greater than δ^k and calculate u_i^k as follows:

$$u_i^k = \frac{1}{\sum_{j=1, s_{ij}^k>\delta^k}^{n} s_{ij}^k}\sum_j \mathbf{I}_{s_{ij}^k>\delta^k} s_{ij}^k w_j^{k-1}. \tag{4}$$

Since the similarity of the client model to itself is always one and larger than δ^k. Like the attention mechanism in deep learning, clients with highly similar data distributions receive higher similarity scores, enabling the server to assign greater weights to their models during the combination [21]. By differentiating between models based on relevance and informativeness, the server can selectively give the most pertinent models to each client rather than treating all models equally. Furthermore, the server can collect weights from clients and utilize the attention mechanism to optimize $\mathcal{R}(W)$ effectively while still ensuring data privacy for all the involved clients.

After optimizing $\mathcal{R}(W)$ on the server, FEDACS then optimizes $\mathcal{F}(W)$ on clients' devices by taking U^k as initialization of W^k and computes w_i^k locally by:

$$w_i^k = u_i^k - \beta_k \nabla \mathbb{F}_i(u_i^k). \tag{5}$$

The update of w_i^k draws inspiration from local fine-tuning in FL that takes the global model as initialization and updates the local models with several gradient steps to fit local data distribution. Considering the scarcity of data, it is more efficient to use u_i^k as an initial value instead of incorporating it into the objective function. This research is driven by prior studies showing difficulties in achieving high training accuracy when the data is either limited or unevenly spread out, thus highlighting the significance of a carefully selected starting point[22].

By iteratively optimizing $\mathcal{F}_\lambda(W)$ through alternating between $\mathcal{R}(W)$ and $\mathcal{F}(W)$ optimization, FEDACS continues this process until a predefined maximum number of iterations, K, is reached.

Algorithm 1: The proposed FEDACS method

Data: n clients, pick ratio p , communication round K

Result: Personalized models $w_1, w_2, ..., w_n$

Initialize clients' models $w_1^0, w_2^0, ..., w_n^0$;
for *round* $k = 1, 2, ..., K$ **do**
  Server randomly picks clients to participate;
  Compute model similarity matrix $\mathcal{S}$;
  Compute p-quantile of model similarities as threshold δ;
  Server updates intermediate models $u_1^k, u_2^k, ..., u_n^k$ for clients by (4);
  Each client i updates local model w_i^k by (5);
end

2.3. Relation to FedAMP

We note that the usage of the regularization term in FEDACS bears a resemblance to FEDAMP[15], a method that also applies attentive message passing to facilitate similar clients to collaborate more. However, in the update steps, FEDACS differs from FEDAMP. First, when computing the intermediate model matrix U, FEDAMP combines all client models for each client. Meanwhile, FEDACS only uses information from clients who share similar data distribution to further reduce the influence of unrelated clients. Meanwhile, if we set the threshold $\delta = \infty$, then the calculation of U is the same as that in FEDAMP. Second, for each round, FEDAMP updates W^k by $W^k = \text{argmin}_W \mathcal{F}(W) + \frac{\lambda}{2\alpha_k}||W - U^k||^2$, which means the W^k needs to reach the local optimal. However, it is not easy to be satisfied when data is insufficient on local devices, and when the learning rate is small enough, the update function in FEDAMP will be simplified to $W^k = U^k$ without using local data information. Additionally, the Euclidean distance version of FEDAMP incorporates an ℓ_2-term regularization, which adds the Euclidean distance of the models to the objective function without using weighted scores. However, this approach provides limited assistance to the server in effectively assigning similar clients' models to each client, thereby underutilizing the potential of the attention mechanism. As discussed above, the limitation of data may adversely influence the performance of FEDAMP, while our proposed FEDACS will not suffer from this issue. The experiment results also demonstrate our conjectures.

2.4. Convergence analysis of FEDACS

In this subsection, we provide the convergence analysis of FEDACS under suitable conditions without the particular requirement of the convexity of the objective function. Consistent with the analysis performed in various incremental and stochastic optimization algorithms [20, 23, 15], we introduce the following assumption and present the main result of the paper.

Assumption 1. There exists a constant $B > 0$ such that $\max\{||Y|| : Y \in \partial\mathcal{F}(W^k)\} \leq B$ and $||\nabla\mathcal{R}(W^k)|| \leq B/\lambda$ hold for every $k \geq 0$, where $\partial\mathcal{F}$ is the subdifferential of $\mathcal{F}$ and $||\cdot||$ is the Frobenius norm.

Theorem 1. Assuming the validity of Assumption 1 and considering the continuous differentiability of functions $\mathcal{R}(W)$ and $\mathcal{F}(W)$, with gradients that exhibit Lipschitz continuity with a modulus L, if $\beta_1 = \beta_2 = \cdots = \beta_K = 1/\sqrt{K}$ and $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \lambda/\sqrt{K}$, then the model matrix sequence $W^0, ..., W^K$ generated by Algorithm 1 satisfies the following.

$$\min_{0\leq k\leq K} ||\nabla\mathcal{F}_\lambda(W^k)||^2 \leq \frac{6(\mathcal{F}_\lambda(W^0) - \mathcal{F}_\lambda^* + 16LB^2)}{\sqrt{K}} + \mathcal{O}\left(\frac{1}{K}\right), \tag{6}$$

where $\mathcal{F}_\lambda^* = \text{argmin}_W \mathcal{F}_\lambda^W$. Additionally, if $\beta_k = \alpha_k/\lambda$ and $\{\alpha_k\}$ satisfies $\sum_{k=1}^\infty \alpha_k = \infty$ and $\sum_{k=1}^\infty \alpha_k^2 < \infty$, then

$$\liminf_{k\to\infty} ||\nabla\mathcal{F}_\lambda(W^k)|| = 0. \tag{7}$$

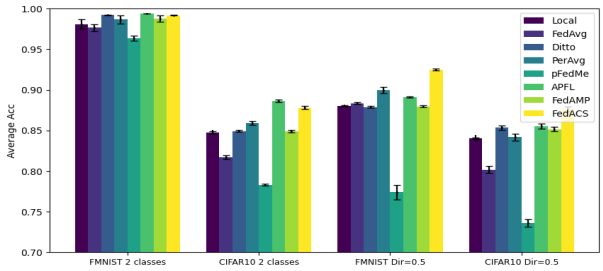
Remark. Theorem 1 implies that for any $\epsilon > 0$, FEDACS needs at most $\mathcal{O}(\epsilon^{-4})$ iterations to find an ϵ-approximate stationary point $\hat{W}$ of problem (1) such that $||\nabla\mathcal{F}_\lambda(\hat{W})|| \leq \epsilon$. It also establishes the global convergence of FEDACS to a stationary point of problem (1) when $\mathcal{F}_\lambda$ is smooth and non-convex. We are unable to provide proof due to page limitations.

3. EXPERIMENTS

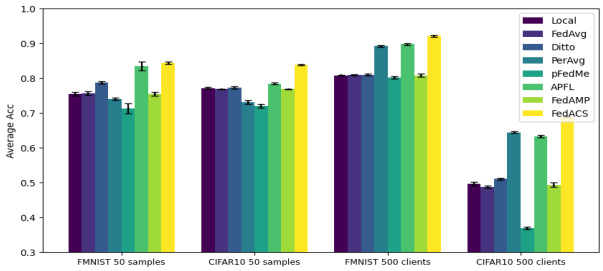
3.1. Experimental settings

We evaluate the performance of FEDACS and compare it with the state-of-the-art PFL algorithms, including DITTO [11], PERAVG [13], PFEDME [12], APFL [9], and FEDAMP [15]. The first three methods are local fine-tuning methods, using the global model to help fine-tune local models. FEDAMP is a method that focuses on adaptively facilitating pairwise collaborations among clients. To ensure the completeness of our experiments, we have also included the results obtained using the standard FL method FEDAVG [2]. The performance of all methods is evaluated by the mean test accuracy across all clients. We perform five experiments for each dataset and partition configuration, wherein we record the mean and variance of the test accuracy. All methods are implemented in PyTorch 1.8 running on NVIDIA 3090.

We use two public benchmark data sets, FMNIST (FashionMNIST) [24], and CIFAR10 [25]. First, we simulate the non-IID and data-sufficient settings by employing two classical data split methods to distribute the data among 100 clients. These methods include: 1) a



(a) Performance on different tasks with sufficient data.



(b) Performance on different tasks with insufficient data.

pathological non-IID data setting, where the dataset is partitioned in a non-IID manner, assigning two classes of samples to each client; and 2) a practical non-IID data setting, where the dataset is partitioned in a non-IID manner following a Dirichlet 0.5 distribution [2]. To simulate scenarios where data insufficiency is encountered during training for each dataset, we adopt two distinct data settings: 1) a pathological non-IID data setting that distributes data into 100 clients following Dirichlet 0.5 distribution, and each user only keeps 50 pieces of data samples for training to simulate the scenario where clients lack sufficient training data: 2) a practical non-IID data setting that distributes data into 500 clients following Dirichlet 0.5 distribution to simulate the scenario of a large number of clients and only some clients participate in each round.

3.2. Results on the sufficient non-IID data setting

Figure 1a presents the mean accuracy and standard deviation per method and sufficient data setting. The two histograms on the left display the performance of PFL methods under pathological settings, while the right-side histograms illustrate the results when the data is partitioned according to the Dirichlet 0.5 distribution. It can be observed that FEDACS performs comparably to most PFL methods but with lower standard deviations. This similarity in performance may be attributed to the availability of sufficient data, allowing the global model to possess adequate generalization ability while its local fine-tuning enables better adaptation to local data distributions. The abundance of training data simplifies the classification task for each client, as evidenced by the strong performance of models trained solely on local datasets. However, non-IID data settings pose challenges for global federated learning methods [15]. FEDAVG's performance on Cifar10 declines noticeably because it introduces instability in the gradient-based optimization process. This instability arises from aggregating personalized models trained on non-IID data from different clients [26]. APFL achieves the best performance in the pathological setting by adaptively learning the model and leveraging the relationship between local and global models, thus increasing the diversity of local models as learning progresses [9]. On the other hand, we can observe that PFEDME struggles to achieve good performance, as it focuses on solving the bilevel problem by aggregating data from multiple clients to improve the global model rather than optimize local performance.

3.3. Results on the insufficient non-IID data setting

Figure 1b presents the mean accuracy, standard deviation per method, and insufficient data setting. The two histograms on the left depict the mean accuracy with only 50 data samples per client, whereas those on the right showcase the results when the data is divided among 500 clients. It is evident that FEDACS outperforms all other methods with smaller standard deviations in both the FMNIST and CIFAR10 datasets. The three local fine-tuning PFL methods face challenges across different datasets and settings, highlighting the fundamental bottleneck in PFL with non-IID data. By relying solely on a misleading global model, the significance of local data distribution in PFL is underestimated, and a single global model fails to capture the intricate pairwise collaboration relationships between clients. It is evident that PERAVG outperforms PFEDME in all settings. Both methods incorporate a "meta-model" during training, but PERAVG utilizes it as an initialization to optimize a one-step gradient update for its personalized model. On the other hand, PFEDME simultaneously pursues both personalized and global models. As discussed in Section 2.3, leveraging the meta-model for a single local update step may yield better results when local devices have limited data. The reason behind PERAVG's suboptimal performance in scenarios where each client possesses only 50 data samples can be attributed to its requirement of three data batches for a single update step. Consequently, the training process becomes insufficient due to the limited data available. As discussed in Section 2.3, FEDAMP fails to leverage the potential of the attention mechanism, resulting in underutilization. Additionally, it requires a large amount of local data to achieve local optima in each round, which is often unrealistic in many federated learning scenarios.

4. CONCLUSION

In this paper, we proposed FEDACS to address the challenge of data heterogeneity and improve overall FL performance. Our approach leveraged the attention aggregation mechanism, allowing clients to identify and collaborate with those who can provide valuable information. This allowed FEDACS to avoid being constrained by a single global model and instead optimize personalized models based on each client's local data distribution. We also analyzed the complexity of FEDACS in achieving first-order optimality. Furthermore, we conducted a series of numerical experiments to demonstrate the superiority of FEDACS in various non-IID settings. The results highlighted its effectiveness in handling data scarcity compared to other methods.

5. ACKNOWLEDGEMENTS

This work was supported in part by the US National Science Foundation (NSF) under awards ECCS-2143559, CNS-2313110, ECCS-2033671, IIS-2006844, IIS-2144209, IIS-2223769, CNS2154962, and BCS-2228534, and the Commonwealth Cyber Initiative under awards VV-1Q23-007, HV-2Q23-003, and VV-1Q24-011.

6. REFERENCES

[1] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[2] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] Zhaohua Zheng, Yize Zhou, Yilong Sun, Zhang Wang, Boyi Liu, and Keqiu Li, "Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges," *Connection Science*, vol. 34, no. 1, pp. 1–28, 2022.

[4] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu, "FFD: A federated learning based method for credit card fraud detection," in *Big Data–BigData 2019: 8th International Congress, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 8*. Springer, 2019, pp. 18–32.

[5] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1–19, 2021.

[6] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[7] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra, "Federated learning with non-IID data," *arXiv preprint arXiv:1806.00582*, 2018.

[8] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.

[9] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi, "Adaptive personalized federated learning," *arXiv preprint arXiv:2003.13461*, 2020.

[10] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant, "Survey of personalization techniques for federated learning," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2020, pp. 794–797.

[11] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6357–6368.

[12] Canh T Dinh, Nguyen Tran, and Josh Nguyen, "Personalized federated learning with Moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21394–21405, 2020.

[13] 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*, vol. 33, pp. 3557–3568, 2020.

[14] Renpu Liu, Jing Yang, and Cong Shen, "Exploiting feature heterogeneity for improved generalization in federated multi-task learning," in *IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 180–185.

[15] 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*, 2021, vol. 35, pp. 7865–7873.

[16] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[17] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[18] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[19] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz, "Deep models under the GAN: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.

[20] Dimitri P Bertsekas et al., "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, vol. 2010, no. 1-38, pp. 3, 2011.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, "On the convergence of FedAvg on non-IID data," *arXiv preprint arXiv:1907.02189*, 2019.

[23] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[25] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[26] Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu, "FedPD: A federated learning framework with adaptivity to non-IID data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.