Federated Learning with GAN-Based Data Synthesis for Non-IID Clients

Zijian Li[1(✉)], Jiawei Shao[1], Yuyi Mao[2], Jessie Hui Wang[3], and Jun Zhang[1]

[1] The Hong Kong University of Science and Technology, Hong Kong, China
zijian.li@connect.ust.hk
[2] The Hong Kong Polytechnic University, Hong Kong, China
[3] Tsinghua University, Beijing, China

Abstract. Federated learning (FL) has recently emerged as a popular privacy-preserving collaborative learning paradigm. However, it suffers from the non-independent and identically distributed (non-IID) data among clients. In this chapter, we propose a novel framework, named *Synthetic Data Aided Federated Learning* (SDA-FL), to resolve this non-IID challenge by sharing synthetic data. Specifically, each client pretrains a local generative adversarial network (GAN) to generate differentially private synthetic data, which are uploaded to the parameter server (PS) to construct a global shared synthetic dataset. To generate confident pseudo labels for the synthetic dataset, we also propose an iterative pseudo labeling mechanism performed by the PS. The assistance of the synthetic dataset with confident pseudo labels significantly alleviates the data heterogeneity among clients, which improves the consistency among local updates and benefits the global aggregation. Extensive experiments evidence that the proposed framework outperforms the baseline methods by a large margin in several benchmark datasets under both the supervised and semi-supervised settings.

Keywords: Federated Learning · Non-Independent and Identically Distributed (non-IID) Problem · Generative Adversarial Network (GAN)

1 Instruction

The recent development of deep learning technologies has led to major breakthroughs in various domains. This results in a tremendous amount of valuable data that can facilitate the training of deep learning models for intelligent applications. A traditional approach to exploit these distributed data samples is to upload them to a centralized server for model training. However, directly offloading data raises severe privacy concerns as data collected from mobile clients may contain sensitive information.

By decoupling model training from the need of transferring private data to the cloud, federated learning (FL) offers a promising approach to collaboratively learn a global model without directly sharing the local data [46]. Particularly, [23] introduced the Federated Averaging (FedAvg) algorithm where the clients train the local models based on the private local data and upload the model updates to the parameter server (PS) for aggregation.

R. Goebel et al. (Eds.): FL 2022, LNAI 13448, pp. 17–32, 2023.
https://doi.org/10.1007/978-3-031-28996-5_2

Despite its success in the independent and identically distributed (IID) scenarios, FL still suffers from significant performance degradation when the data distribution among clients becomes skewed. In particular, different clients learn from different data distributions in the non-IID scenarios, which leads to high inconsistency among the local updates and thus degrades the effectiveness of global model aggregation [34].

Many works have been proposed to alleviate the non-IID issue by regularizing the local models with the knowledge of the global model and local models from other clients [1,13,18,34]. These methods, however, aim to reduce the local model bias and cannot achieve a significant improvement in scenarios with extreme non-IIDness [17]. Recent studies have also attempted to tackle the non-IID problem with data augmentation techniques [46]. Specifically, [27,32,39] proposed to generate synthetic samples by mixing the real samples. Nevertheless, without implementing a privacy-protection mechanism, these methods are susceptible to data leakage. In addition, another recent work attempts to overcome the non-IID issue via secrete data sharing [28], but it confronts with additional communication costs and implementation challenges.

Observing the data heterogeneity problem and the privacy leakage of the existing data augmentation methods for FL, we propose a novel framework, named *Synthetic Data Aided Federated Learning* (SDA-FL), which resolves the non-IID issue by sharing the differentially private synthetic data. In this framework, each client pretrains a local differentially private generative adversarial network (GAN) [8] to generate synthetic data, thus avoiding sharing the raw data. These synthetic data are then collected by the PS to construct a global synthetic dataset. To generate confident pseudo labels for the synthetic data, we propose an iterative pseudo label update mechanism, in which the PS utilizes the received local models to update the pseudo labels in each training round. As the local models are progressively improved over the FL process, the confidence of pseudo labels is thus enhanced, which is beneficial for the server updates and local updates in future rounds and in turn results in a well-performed global model. It is worth noting that the SDA-FL framework is compatible with many existing FL methods and can be applied in both supervised and semi-supervised settings without requiring labels of the real data, which will be validated in the experiments. Ablation studies are also conducted to illustrate the impact of the privacy budget and the effectiveness of the key procedures in SDA-FL.

2 Related Works

Non-IID Challenges in Federated Learning. The non-IID data distribution has been a fundamental obstacle for FL [46]. This is because the highly skewed data distribution significantly amplifies the local model divergence and thus deteriorates the performance of the aggregated model [20,44]. To mitigate the client drift caused by the non-IID data, many works proposed to modify the local objective function with the additional knowledge from the global model and local models of other clients [1,13,18,34]. Such methods, however, cannot achieve satisfactory performance in many non-IID scenarios [17]. In addition to training the same model structure at clients, some studies proposed to combat the negative impact of non-IID data by keeping the specific local model structures individually, including local batch normalization layers [19], local extractors [2,21], and local classifiers [47]. Moreover, some prior researches addressed the data heterogeneity problem by optimizing the operations at the PS, such as model aggregation [34], client selection [33,43], client clustering [7,14], classifier calibration [20,22,45], and domain adaptation [31,36].

Data Augmentation and Privacy Preserving. Recently, FL methods based on some form of data sharing have received increasing attention for their prominent performance [44,46]. A popular approach is to leverage the Mixup technique [41] for data augmentation, so that the clients can share the blended local data and collaboratively construct a new global dataset to tackle the non-IID issue [27,29,39]. However, simply combination of the real samples may be vulnerable to privacy attacks. Alternatively, GAN-based data augmentation [11,40,44] was shown to be effective in reducing the degree of local data imbalance in FL. The general idea is to train a good generative model at the server based on a few seed data samples uploaded by the clients. Then this well-trained generator is downloaded by all clients for local model updating. Nevertheless, since sending local data samples to the server violates the data privacy requirement, FedDPGAN [42] suggested all the clients collaboratively train a global generative model based on the FL framework to supplement the scarce local data. Unfortunately, the GAN training process also requires frequent generative models exchanges, leading to extremely high communication costs and risks of adversarial attacks [4]. In addition, existing GAN-based methods require the fully labeled data at clients to train the supervised GANs to generate the synthetic samples with labels, which is impractical in Internet of Things (IoT) and healthcare systems. Instead, our proposed algorithm allow clients to train the unsupervised GANs locally without requiring the labeled data and it is able to provide confident pseudo labels for the synthetic samples.

3 Preliminary

Federated Learning. Federated learning aims to train a promising global model $\boldsymbol{w}$ without disclosing any local data samples of clients. In each training round $t = 0, 1, \ldots, T-1$, FedAvg [23] tries to minimize the global objective function $F(\boldsymbol{w}_t)$ as follows:

$$\min_{\boldsymbol{w}_t} F(\boldsymbol{w}_t) \triangleq \sum_{k \in \mathcal{S}_t} p_k F_k(\boldsymbol{w}_t), \quad p_k = \frac{|\mathcal{D}_k|}{\sum_{i \in \mathcal{S}_t} |\mathcal{D}_i|}, \tag{1}$$

where $\mathcal{S}_t$ is the subset of clients activated in round t, p_k is the aggregation weight for client k that is normally chosen according to the size of its local dataset $\mathcal{D}_k$, and $F_k(\boldsymbol{w}_t)$ is the local objective function of client k in round t defined as follows:

$$F_k(\boldsymbol{w}_t) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_k} \ell(\boldsymbol{w}_{t-1}^k; x, y), \tag{2}$$

where $\ell(\cdot)$ is the cross-entropy loss and $\boldsymbol{w}_{t-1}^k$ is the global model downloaded from the PS. Specifically, in each local training step $e = 0, 1, \ldots, E-1$, every client in set $\mathcal{S}_t$ updates the local model with the real batch samples $(\mathbf{X}_{t,e}^k, \mathbf{Y}_{t,e}^k)$ via stochastic gradient descent (SGD), i.e., $\boldsymbol{w}_{t,e+1}^k \leftarrow \boldsymbol{w}_{t,e}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,e}^k; \mathbf{X}_{t,e}^k, \mathbf{Y}_{t,e}^k)$, $\boldsymbol{w}_t^k \leftarrow \boldsymbol{w}_{t,E-1}^k$. The updated local models $\boldsymbol{w}_t^k$ are then sent back to the PS for weighted aggregation. These procedures repeat until all the T training rounds are exhausted.

With the iterative local training and global aggregation procedures, the PS is expected to obtain a well-performed global model even without access to any local data. However, the highly skewed data distribution among clients easily leads to severe local model divergence and consequently degrades the global model performance [17]. To resolve this issue and avoid sharing the real data, we exploit the generative adversarial network (GAN) to generate high-quality synthetic data that can be shared among clients, which are used to update the local and global models.

Differentially Private Generative Adversarial Network. To avoid the gradient vanishing and mode collapse problems encountered by conventional GAN models [3], the Wasserstein GAN with gradient penalty (WGAN-GP) [9], which penalizes the gradient norm of the critic to stabilize the training process of the generator G and discriminator D, is adopted. With the real data distribution $p_r(x)$ and input noise distribution of the generator $p_z(z)$, the objective function of the WGAN-GP is expressed as follows:

$$\min_G \max_D \mathbb{E}_{x\sim p_r(x)}\left[D(x)\right] - \mathbb{E}_{z\sim p_z(z)}\left[D\left(G(z)\right)\right] + \gamma \left\|\nabla_{\hat{x}} D(\hat{x}) - 1\right\|_2^2, \tag{3}$$

where $\hat{x}$ is a mixture of the real sample x and the fake sample $G(z)$, and γ is a hyperparameter. To provide differential privacy protection for the synthetic data, we inject Gaussian noise into the GAN training process. The definition of differential privacy (DP) is given as follows:

Definition 1. *(Differential privacy [6]): A random mechanism $\mathcal{A}_p$ satisfies (ϵ, δ)-differential privacy if for any output's subset ($\mathcal{S}$) and any two adjacent datasets $\mathcal{M}$, $\mathcal{M}'$, the following probability inequality holds:*

$$\mathbb{P}\left(\mathcal{A}_p(\mathcal{M}) \in \mathcal{S}\right) \leq e^{\epsilon} \cdot \mathbb{P}\left(\mathcal{A}_p\left(\mathcal{M}'\right) \in \mathcal{S}\right) + \delta, \tag{4}$$

where $\delta > 0$ and ϵ is the privacy budget indicating the privacy level, i.e., a smaller value of ϵ implies stronger privacy protection.

To satisfy the (ϵ, δ)-DP, we follow [38] and add Gaussian noise to the updated gradients at each discriminator training iteration. The relationship between the noise variance and differential privacy is shown below:

$$\sigma_n = \frac{2q}{\epsilon}\sqrt{n_d \log\left(\frac{1}{\delta}\right)}, \tag{5}$$

where q and n_d denote the sample probability for each instance and the total batch number of the local dataset, respectively. Besides, according to the post-processing property of DP [6], any mapping from the differentially private output also satisfies the same level of DP. In other words, the gradients of the generator, which are obtained via the backpropagation from the noisy discriminator output, also meet the (ϵ, δ)-DP condition.

Based on the training algorithm of WGAN-GP, it is desirable to train the generators with the label information, e.g., Auxiliary Classifier Generative Adversarial Network (ACGAN) [25,26], so that the synthetic data can be generated with labels. However, considering the label scarcity problem in the federated semi-supervised settings, this form of the conditional generative model cannot be trained at clients. Therefore, we resort to a more general paradigm that trains an unsupervised generator, and propose a pseudo labeling procedure within the FL process to generate high-confidence pseudo labels.

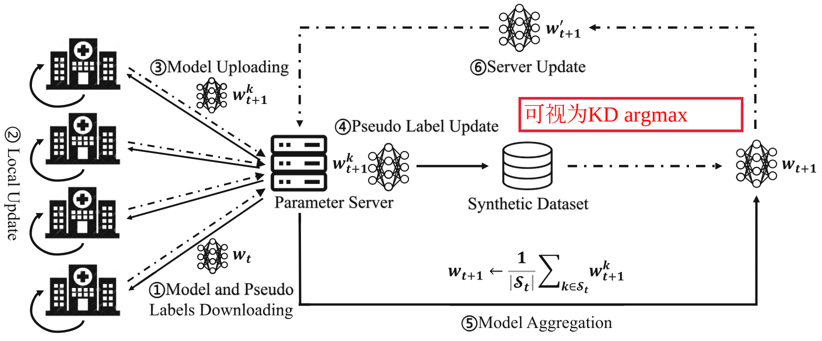
Pseudo Labeling. To generate confident pseudo labels, following [30], only the class with an extremely high prediction probability is regarded as the pseudo label. Specifically, with a predefined threshold τ, class c is deemed as the label of sample x if the output prediction probability $f_c(\boldsymbol{w};x)$ is the largest among all the classes and also larger than the threshold τ. Hence, the pseudo labeling function can be expressed as follows:

$$\hat{y} = \begin{cases} c & \text{if } \max_c f_c(\boldsymbol{w};x) > \tau, \\ \text{None} & \text{otherwise.} \end{cases} \tag{6}$$

超过预定阈值则
分配pesudo label

With such a pseudo labeling procedure, only the high-quality synthetic data can output a high prediction probability by model $\boldsymbol{w}$ and obtain their pseudo labels. As such, the unqualified synthetic samples are filtered, leaving only the qualified ones to update the local models.

In our proposed FL framework, the local models are used to predict the pseudo labels for the synthetic data generated by the corresponding local generators, and the pseudo labels are continuously updated with the improved local models during the FL process, as will be discussed in the next section.



可视为KD argmax

Fig. 1. Overview of the proposed SDA-FL framework. Before the FL process starts, the synthetic data from all clients are sent to the PS to construct a global synthetic dataset. In each training round, every client first downloads the global model and updates the pseudo labels of the synthetic data for local training. The local models are then uploaded to the PS for pseudo label updating and model aggregation. Lastly, the PS updates the global model $\boldsymbol{w}_{t+1}$ with the updated synthetic dataset.

4 Synthetic Data Aided Federated Learning (SDA-FL)

We now introduce the SDA-FL framework that adopts GAN-based data augmentation to alleviate the negative effect of the non-IID data. The overview of the SDA-FL framework is shown in Fig. 1, and key algorithmic innovations built upon the classic FL framework are elaborated below.

Global Synthetic Dataset Construction. At the start of the FL process, we propose to train the generative model based on the local samples and then utilize it to generate synthetic samples to tackle the non-IID problem. Although there are some recent works suggest to collaboratively train the global generator based on the FL framework [25,42], such a cooperative training strategy for GANs requires significant communication bandwidth and incurs additional privacy leakage due to frequent exchanges of the generators and discriminators.

1, client预训练阶段生成数据并上传至server

To avoid these problems, we resort to local training strategies for the generative models. In particular, each client pretrains a local GAN model to generate synthetic samples based on its local data. Then, the synthetic samples are sent to the PS to construct a global shared synthetic dataset. To effectively leverage the synthetic dataset for FL, we perform pseudo labeling for these samples, which is critical to the effectiveness of the SDA-FL framework.

Synthetic Samples Annotation. Unlike the prior work in [12] where each client leverages the local models from other clients to annotate the unlabeled data that encounters a bottleneck with highly skewed data distribution, we only utilize the local models to perform pseudo labeling for the corresponding unlabeled synthetic data. This is because the local model and the corresponding synthetic data are trained with the same local data at each client, and only the high-quality synthetic samples can obtain a high prediction probability with this local model. In addition, the confidence of the pseudo labels heavily relies on the local model quality, but the under-trained local models at the beginning of the FL process fail to accomplish this task. Therefore, we update the pseudo labels for the global shared synthetic dataset with the improved local models in each FL round. Specifically, after receiving the local model $\boldsymbol{w}_t^k$ in round t, the PS assigns a pseudo label for each unlabeled synthetic instance x according to (6), i.e., its maximum class probability $f_c(\boldsymbol{w}_t^k; x)$ is higher than the predefined threshold τ. In this way, we can gradually generate high-quality pseudo labels for the synthetic data samples.

2, 基于synthetic data对应的client local model更新Label

Synthetic Data Aided Model Training. Augmented by the samples $\hat{\mathbf{X}}$ and confident pseudo labels $\hat{\mathbf{Y}}_t$ from the shared synthetic dataset, the data available for local training at different clients are approximately homogeneously distributed. To make good use of the synthetic data, we leverage the Mixup method proposed in [41], which utilizes a linear interpolation between the real batch samples $(\mathbf{X}_{t,e}^k, \mathbf{Y}_{t,e}^k)$ and the synthetic batch samples $(\hat{\mathbf{X}}_e, \hat{\mathbf{Y}}_{t,e})$, to augment the real data for client k at the local training step e of round t:

$$\begin{aligned} \bar{\mathbf{X}}_{t,e}^k &= \lambda_1 \hat{\mathbf{X}}_e + (1-\lambda_1)\mathbf{X}_{t,e}^k, \\ \bar{\mathbf{Y}}_{t,e}^k &= \lambda_1 \hat{\mathbf{Y}}_{t,e} + (1-\lambda_1)\mathbf{Y}_{t,e}^k, \end{aligned} \tag{7}$$

3, 用mixup策略增强数据，这里应该是在client 执行

where λ_1 follows the Beta distribution for each batch, i.e., $\mathrm{Beta}(\alpha, \alpha)$ with $\alpha \in [0, 1]$. By combining the cross-entropy loss $\ell(\cdot)$, the mixup loss for the local model update becomes:

$$\ell_1 = \lambda_1 \ell\big(f(\bar{\mathbf{X}}^k_{t,e}; \boldsymbol{w}^k_{t,e}), \hat{\mathbf{Y}}_{t,e}\big) + (1 - \lambda_1)\ell\big(f(\bar{\mathbf{X}}^k_{t,e}; \boldsymbol{w}^k_{t,e}), \mathbf{Y}^k_{t,e}\big). \tag{8}$$

In addition, since the loss in (8) is fragile at the beginning of the FL process caused by the unconfident pseudo labels, another cross-entropy loss term is introduced for the real batch samples $(\mathbf{X}^k_{t,e}, \mathbf{Y}^k_{t,e})$ to stabilize the training process:

$$\ell_2 = \ell\big(f(\mathbf{X}^k_{t,e}; \boldsymbol{w}^k_{t,e}), \mathbf{Y}^k_{t,e}\big). \tag{9}$$

Then, SGD is applied to update the local model as follows:

$$\boldsymbol{w}^k_{t,e+1} \leftarrow \boldsymbol{w}^k_{t,e} - \eta_t \nabla(\ell_1 + \lambda_2 \ell_2), \qquad \boxed{\text{Local client training}} \tag{10}$$

where λ_2 is a hyper-parameter to control the retention of the local data.

In contrast to traditional FL where the PS does not have access to any data to update the global model, the PS in our framework keeps the entire global synthetic dataset $\hat{\mathcal{D}}_s$ and uses it to train the global model. Particularly, since there is no real data in the PS, two batches of synthetic samples are used to update the global model with (10) at each iteration.

Interplay Between Model Training and Synthetic Dataset Updating. In each FL round, the aid of synthetic data improves the generalization of local models. Since the updated local models are used for pseudo labeling and the global synthetic dataset construction at the PS, the confidence of the pseudo label is thus boosted. With the enhanced synthetic dataset, the PS can refine the global model and all the clients can improve their local models subsequently at the next round. Therefore, the interplay between model training and synthetic dataset updating at every training round is critical to achieving a well-performed global model.

SDA-FL vs. Traditional FL. Compared with traditional FL, the proposed SDA-FL framework introduces additional operations at both the clients and PS. These innovations contribute to the performance improvement of FL with non-IID data. Specifically, in traditional FL algorithms [13, 18], clients update their models based only on the local data, which easily leads to performance degradation when data are heterogeneous among clients. In our framework, the local datasets are augmented by the GAN-based synthetic samples to alleviate the non-IID problem. Furthermore, the PS in traditional FL algorithms only performs simple model aggregation. In contrast, the PS in the SDA-FL updates the global model with the high-confidence synthetic data, which further improves the global model performance. Overall, with a shared synthetic dataset and an iterative pseudo labeling mechanism, SDA-FL overcomes the issue of heterogeneous data distributions among clients and enhances the global model at the PS. We envision that this framework can be extended to develop other data augmentation-based methods for both federated supervised learning and federated semi-supervised learning.

5 Experiments

In this section, we evaluate the proposed SDA-FL framework in the presence of non-IID data for both federated supervised learning and federated semi-supervised learning. The experimental results on different benchmark datasets demonstrate the superiority of the proposed framework over the baseline methods. Ablation studies are also conducted to discuss the effectiveness of key procedures and hyper-parameters in SDA-FL.

5.1 Experiment Setup

Datasets. Following [17], we use four benchmark datasets, including MNIST [16], FashionMNIST [37], CIFAR-10 [15], and SVHN [24], to evaluate the proposed method. In all the experiments, we equally divide the training samples and assign them to the clients. Specifically, given the number of classes per client as C and total K clients, the whole training dataset is split into $K \times C$ subsets, and each subset only has a single class. Then all the subsets of data are randomly shuffled and distributed to the clients. Besides, a Dirichlet-based label imbalance distribution is also used to validate all the methods, for which we sample $p_c \sim Dir_K(0.1)$ and allocate a proportion $p_{c,k}$ of samples of class c to client k. We assume two classes of FashionMNIST data at each client in the ablation studies.

To guarantee the effectiveness of SDA-FL, we set the hyper-parameters γ, λ_2, and threshold τ to be 10.0, 1.0, and 0.95, respectively. We deploy ten clients in the experiments, and all of them are selected in each round. To generate high-quality synthetic data, we pretrain the generator and discriminator with 36,000 iterations for CIFAR-10 and 18,000 for the other datasets in both the federated supervised learning experiments and federated semi-supervised learning experiments. Each client uploads 4,000 synthetic samples to the PS to construct the global synthetic dataset. There are 200 rounds for all the methods. In each round of SDA-FL, the PS utilizes the synthetic dataset to update the global model with 10 iterations for CIFAR-10 and 50 iterations for the other datasets. Besides, we set the local training step size $E = 90$ in the federated supervised learning experiments and $E = 40$ in the federated semi-supervised learning experiments, and select the SGD with learning rate $\alpha = 0.03$ as the optimizer. In each iteration, the clients update the local models with batch size $B = 64$ for the federated supervised learning experiments and $B = 80$, including 16 labeled samples and 64 unlabeled samples, for the federated semi-supervised learning experiments.

Moreover, to evaluate the proposed SDA-FL in practical applications, we also test all the methods on a realistic COVID-19 dataset [35]. Because of the scarcity of Pneumonia samples, we only assume six clients in this experiment, each of which has two classes of data as shown in Table 1. We train the GAN models locally with 4,500 iterations, and update the local models with 30 iterations and the global model with 10 iterations.

Table 1. Data distribution of the COVID-19 dataset.

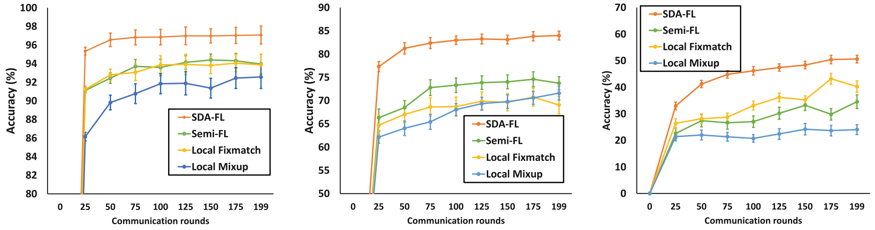
Hospital	**0**	**1**	**2**	**3**	**4**	**5**
Data	Normal 2,000 COVID-19 750	COVID-19 750 Pneumonia 250	Pneumonia 250 Normal 2,000	Normal 2,000 COVID-19 750	COVID-19 750 Pneumonia 250	Pneumonia 250 Normal 2,000

Table 2. Test accuracy (%) of different methods on various datasets. $\#C$ represents the number of sample classes at each client.

Datasets	Data distribution	FedAvg	FedProx	SCAFFOLD	Naivemix	FedMix	SDA-FL
MNIST	$p_k \sim Dir(0.1)$	96.71	96.78	96.72	96.03	96.55	**98.20**
	$\#C = 1$	83.44	84.17	25.39	84.35	90.96	**98.19**
	$\#C = 2$	97.61	97.55	94.17	84.35	90.96	**98.26**
	$\#C = 3$	98.42	98.38	96.89	98.11	98.46	**98.50**
FashionMNIST	$p_k \sim Dir(0.1)$	79.76	80.02	80.10	78.68	79.02	**91.20**
	$\#C = 1$	16.50	57.14	56.80	66.62	72.11	**85.70**
	$\#C = 2$	73.50	75.76	70.82	79.54	82.41	**86.87**
	$\#C = 3$	82.47	83.43	77.68	82.09	84.65	**87.06**
CIFAR-10	$p_k \sim Dir(0.1)$	78.20	77.71	77.57	75.92	77.92	**82.57**
	$\#C = 1$	18.36	11.24	12.81	14.39	13.57	**37.70**
	$\#C = 2$	61.28	63.16	60.78	64.39	65.76	**67.89**
	$\#C = 3$	79.33	79.54	79.35	78.92	79.49	**84.56**
SVHN	$p_k \sim Dir(0.1)$	90.22	90.32	90.01	90.05	91.24	**92.41**
	$\#C = 1$	14.05	17.53	11.64	14.35	16.78	**88.46**
	$\#C = 2$	81.11	86.28	73.34	84.64	86.61	**90.70**
	$\#C = 3$	84.18	92.15	80.13	92.30	92.61	**93.16**

Baselines. For the federated supervised learning experiments, we compare the SDA-FL framework with FedAvg [23], FedProx [18], SCAFFOLD [13], Naivemix, and FedMix [39] on the MNIST, FashionMNIST, CIFAR-10, SVHN, and COVID-19 datasets. For the COVID-19 dataset, we also adopt FedDPGAN [42] for comparisons, which trains a global GAN to resolve the non-IID issue for medical applications. We report the best results by tuning the hyperparameter μ of the regularization term for FedProx and the mixup ratio λ for FedMix. Besides, we extend our framework to the semi-supervised learning setting on the MNIST, FashionMNIST, and CIFAR-10 datasets by performing pseudo labeling for the unlabeled local data. We compare the SDA-FL framework with Semi-FL [5], Local Fixmatch [30], and Local Mixup [41] to show its effectiveness.

Models. We adopt a simple CNN model that consists of two convolutional layers and two fully-connected layers for the MNIST and FashionMNIST classification tasks. Meanwhile, ResNet18 [10] is used for classifying the CIFAR-10, SVHN, and COVID-19 datasets. To generate qualified synthetic samples, we use a generator with four deconvolution layers and a discriminator with four convolutional layers followed by a fully-connected layer.



(a) MNIST (b) FashionMNIST (c) CIFAR-10

Fig. 2. Test accuracy of different methods for federated semi-supervised learning on the MNIST, FashionMNIST, and CIFAR-10 datasets.

5.2 Evaluation Results

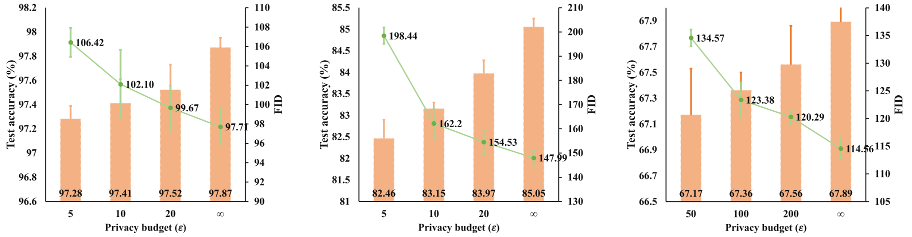
Performance in Federated Supervised Learning. With varying numbers of classes per client, the experimental results in Table 2 show that our framework outperforms the baselines by a significant margin, which attributes to the GAN-based data augmentation that mitigates the detrimental effects of the data heterogeneity on FL. Under the severe non-IID scenario (i.e., each client only has one class of data), our SDA-FL method maintains an accuracy of 88.46% in the SVHN classification task, while other baselines suffer from great performance degradation. In the CIFAR-10 experiments, our framework is superior to the Naivemix and FedMix algorithms at least by 5.0% with three classes of data at each client, which verifies the competence of the GAN-based data compared with the mixing data.

In the COVID-19 experiments, besides the better performance over the aforementioned baselines, SDA-FL also surpasses FedDPGAN by 1.68% in accuracy. This demonstrates that the individually trained GANs generate synthetic data of higher quality than the global GAN trained based on the FL framework. Furthermore, in addition to resolving the non-IID issue, we find that SDA-FL even outperforms FedAvg (IID) by 1.14% in accuracy, which shows its advantages in supplementing more valuable training samples for the local training and thus improving the generalizability of the global model (Table 3).

Table 3. Test accuracy (%) on the COVID-19 dataset. FedAvg (IID) represents the scenario where the training samples are uniformly distributed to all clients to achieve the IID distribution.

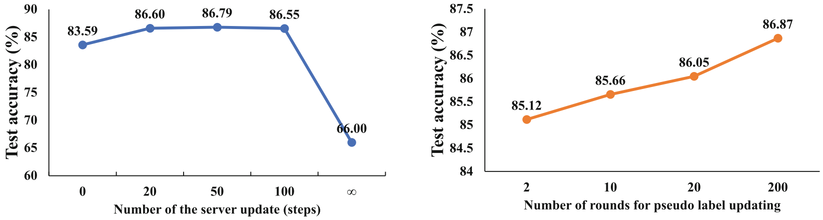
Algorithm	FedAvg	FedProx	SCAFFOLD	Naivemix	FedMix	FedDPGAN	FedAvg (IID)	SDA-FL
Accuracy	94.05	95.03	94.30	94.14	94.28	94.57	95.19	**96.25**

Performance in Federated Semi-Supervised Learning. The results in Fig. 2 show that the SDA-FL framework achieves faster convergence and better performance than other algorithms in the federated semi-supervised learning setting, indicating its robustness and generalizability. Particularly, compared with Semi-FL, our method improves



(a) MNIST (b) FashionMNIST (c) CIFAR-10

Fig. 3. Test accuracy and FID score with respect to the privacy budget. We run three trails and report the mean and the standard deviation of the test accuracy. The FID scores of the real samples on MNIST, FashionMNIST, and CIFAR-10 are 10.54, 23.17, and 42.70, respectively, which are much larger than those of the synthetic data.



(a) Server update (b) Pseudo label update

Fig. 4. Test accuracy on FashionMNIST with varying step sizes for server updating and rounds for pseudo label updating. The "∞ steps" in (a) means that the model is only trained with the synthetic data, and the "10 rounds" in (b) represents that the PS only updates the pseudo labels in the first 10 rounds.

the accuracy by almost 10% on the FashionMNIST classification task. In the CIFAR-10 dataset, the baseline methods are not able to train a usable global model (i.e., with a test accuracy below 40%), while the proposed framework converges in this challenging scenario and improves the test accuracy significantly. This is because the proposed pseudo labeling mechanism can provide high-quality labels for the synthetic and unlabeled samples, which are beneficial to the FL process.

Tradeoff Between the Privacy Budgets and Model Performance. To investigate the impact of the privacy budgets, we evaluate the model performance of the SDA-FL framework under different values of ϵ. The Fréchet inception distance (FID) is used to measure the quality of the generated samples, where a smaller FID score indicates better image quality. As illustrated in Fig. 3, a strict privacy budget of $\epsilon = 5$ increases the FID score compared with that in the protection-free scenario, which implies quality degradation of the generated samples. This negative impact on the synthetic samples also leads to around 0.61% and 2.59% accuracy drop on the MNIST and FashionMNIST datasets, respectively. The CIFAR-10 classification task follows a similar trajectory. Please note that although the proposed SDA-FL framework is trained under strict privacy requirements, compared with the results in Table 2, it still maintains supreme performance.

Table 4. Test accuracy and FID comparison with WGAN-GP/AC-WGAN-GP on various datasets.

Datasets	FashionMNIST		CIFAR-10 (2class/client)		CIFAR-10 (3class/client)	
Algorithms	WGAN-GP	AC-WGAN-GP	WGAN-GP	AC-WGAN-GP	WGAN-GP	AC-WGAN-GP
FID	**217.81**	220.39	**114.56**	154.27	**129.25**	162.16
Accuracy (%)	**83.76**	82.03	**67.89**	67.22	**84.56**	83.53

Effectiveness of Server Update and Pseudo Label Update. In comparison to the traditional FL, our framework updates the global model with the synthetic data at the PS, which has the potential to further improve the performance. The results in Fig. 4(a) show that the model performance on FashionMNIST reduces by nearly 3% without any server updates. Nevertheless, updating the global model too much by the PS may degrade the performance because of the excessive involvement of synthetic data. Empirical results show that the model trained solely with the synthetic data (i.e., the server updates the global model for ∞ steps) can only obtain an accuracy of 66.0%, which highlights the necessity of judicious utilization of the synthetic and local data for model training.

Besides, keeping updating pseudo labels in each round adopted by the SDA-FL framework for the synthetic data improves the model performance. As illustrated in Fig. 4(b), the accuracy increases with the number of rounds for pseudo label updating, which demonstrates that the SDA-FL framework can improve the confidence level of the pseudo labels over the training process. Note that since our framework only transmits the pseudo labels instead of the synthetic samples, the extra communication overhead is negligible.

Performance Comparison with Auxiliary Classifier WGAN-GP (AC-WGAN-GP). We can also include the label information in the GAN training to generate synthetic data with labels in the federated supervised learning experiments. As such, we compare the performance of SDA-FL with WGAN-GP and AC-WGAN-GP [26] on the FashionMNIST and CIFAR-10 datasets. As shown in Table 4, with the same number of training iterations for the generators, WGAN-GP achieves higher synthetic data quality as implied by the higher FID scores. Although AC-WGAN-GP can generate labeled synthetic data, WGAN-GP still performs better in accuracy performance with the higher-quality synthetic data. This is because our proposed pseudo labeling mechanism provides confident pseudo labels for the synthetic data.

Ablation Studies on the Computational Cost and Number of Samples for Training the Generators. We present the results in Fig. 5 and Fig. 6, which show that insufficient training rounds and training samples for generators result in low-quality synthetic data and degraded performance. The generator only gets an FID score of 217.81 (23.17 for real data) and an accuracy of 83.76% when training the generators with 30 rounds. In addition, as shown in Fig. 6, under the fixed 30-round computational cost, using fewer samples for training the generators reduces the quality of the synthetic data and thus affects the performance. Nonetheless, despite the low computational cost (30 rounds) and the small number of samples (1000 samples) used to train the generators individually, our framework still outperforms other baselines (82.41% in Fedmix), demonstrating the robustness of the proposed method.

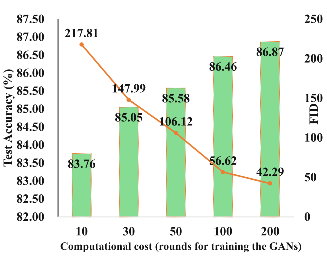


Fig. 5. Test accuracy and FID on Fashion-MNIST, under varying training rounds (i.e., computational cost) for the GANs.

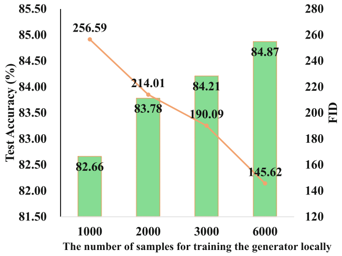


Fig. 6. Test accuracy and FID on FashionMNIST, under varying samples for training the GANs locally.

Ablation Study on the Size of the Synthetic Dataset. Before the federated learning process, all clients construct the synthetic dataset together. Larger synthetic datasets can typically provide more qualified data to clients and PS to improve the performance, but they require larger storage space for clients. The results in Fig. 7 show that, although the test accuracy on FashionMNIST decreases as the size of the synthetic dataset declines, our framework achieves 85.44% test accuracy with each client only uploading 200 samples, which is acceptable in exchange for less storage space requirement.

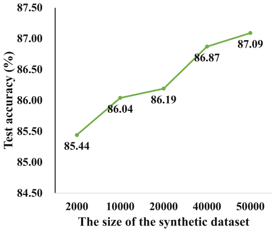


Fig. 7. Test accuracy on FashionMNIST, under varying sizes of the synthetic dataset.

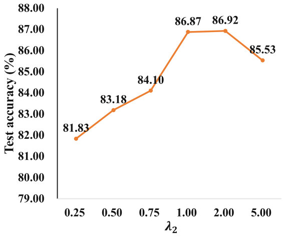


Fig. 8. Test accuracy on FashionMNIST, under varying values of λ_2 in the local update.

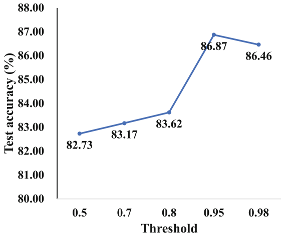


Fig. 9. Test accuracy on FashionMNIST, under varying threshold values.

Ablation Study on the Hyper-parameter $\boldsymbol{\lambda_2}$. λ_2 is the parameter to control the use of the local real data. The results in Fig. 8 show that compared with other values of λ_2, the test accuracy on FashionMNIST drops by nearly 5.0% with $\lambda_2 = 0.25$, indicating that the training of the models is heavily reliant on the real data. Besides, an excessive value of λ_2 also degrades the performance since it eliminates the benefits of the synthetic data and cannot mitigate the non-IID problem as much as possible. In comparison, a moderate value of λ_2 is capable of combining the real data and the synthetic data intelligently and thus enhancing the models and solving the non-IID issue in FL.

Ablation Study on the Threshold τ. The threshold τ sets the criterion for selecting the high-confidence pseudo labels. The results on FashionMNIST are shown in Fig. 9, which indicates that a small threshold impairs the performance because some synthetic data with low-confidence pseudo labels are still considered qualified and used to train the models. A too large threshold value, on the other hand, filters out a lot of qualified data, which affects performance. In our experiment setup, $\tau = 0.95$ is a reasonable value to strike a good balance between quality and the quantity of the pseudo labels.

6 Conclusions and Discussions

We proposed a new data augmentation method to resolve the heterogeneous data distribution problem in federated learning by sharing the differentially private GAN-based synthetic data. To effectively utilize the synthetic data, a novel framework, named Synthetic Data Aided Federated Learning (SDA-FL), was developed, which generates and updates confident pseudo labels for the synthetic data samples. Experiment results showed that SDA-FL outperforms many existing baselines by remarkable margins in both supervised learning and semi-supervised learning under strict differential privacy protection. In this study, we limit our attention to the WGAN-GP and AC-WGAN-GP in SDA-FL. Despite their performance improvements compared with the baselines, it is interesting to investigate other GAN structures. In addition, WGAN-GP requires significant computational resources at clients. Therefore, to improve the applicability of SDA-FL, it is important to develop a computation-efficient GAN-based structure for clients in future research.

References

1. Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: International Conference on Learning Representations (2020)
2. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017)
4. Chen, D., Yu, N., Zhang, Y., Fritz, M.: GAN-leaks: a taxonomy of membership inference attacks against generative models. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 343–362 (2020)
5. Diao, E., Ding, J., Tarokh, V.: Semifl: Communication efficient semi-supervised federated learning with unlabeled clients. arXiv e-prints, p. arXiv-2106 (2021)
6. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
7. Ghosh, A., Chung, J., Yin, D., Ramchandran, K.: An efficient framework for clustered federated learning. In: Conference on Neural Information Processing Systems (NIPS) (2020)
8. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: NIPS (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

11. Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.L.: Communication-efficient on-device machine learning: federated distillation and augmentation under non-IID private data. arXiv preprint arXiv:1811.11479 (2018)
12. Jeong, W., Yoon, J., Yang, E., Hwang, S.J.: Federated semi-supervised learning with inter-client consistency & disjoint learning. In: International Conference on Learning Representations (2020)
13. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: SCAFFOLD: stochastic controlled averaging for federated learning. In: ICML (2020)
14. Kopparapu, K., Lin, E., Zhao, J.: FedCD: improving performance in non-IID federated learning. arXiv preprint arXiv:2006.09637 (2020)
15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Master's thesis, University of Tront (2009)
16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-IID data silos: an experimental study (2021). https://doi.org/10.48550/ARXIV.2102.02079, https://arxiv.org/abs/2102.02079
18. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLSys) (2020)
19. Li, X., JIANG, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: federated learning on non-IID features via local batch normalization. In: International Conference on Learning Representations (2020)
20. Li, X.C., Zhan, D.C.: FedRS: federated learning with restricted softmax for label distribution non-IID data. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 995–1005 (2021)
21. Liang, P.P., et al.: Think locally, act globally: federated learning with local and global representations. arXiv preprint arXiv:2001.01523 (2020)
22. Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., Feng, J.: No fear of heterogeneity: classifier calibration for federated learning with non-IID data. Adv. Neural Inf. Process. Syst. **34** (2021)
23. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2017)
24. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
25. Nguyen, D.C., Ding, M., Pathirana, P.N., Seneviratne, A., Zomaya, A.Y.: Federated learning for COVID-19 detection with generative adversarial networks in edge cloud computing. IEEE Internet Things J. **9**, 0257–10271 (2021)
26. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: International Conference on Machine Learning, pp. 2642–2651. PMLR (2017)
27. Oh, S., Park, J., Jeong, E., Kim, H., Bennis, M., Kim, S.L.: Mix2fld: downlink federated learning after uplink federated distillation with two-way mixup. IEEE Commun. Lett. **24**, 2211–2215 (2020)
28. Shao, J., Sun, Y., Li, S., Zhang, J.: DReS-FL: dropout-resilient secure federated learning for non-IID clients via secret data sharing (2022)
29. Shin, M., Hwang, C., Kim, J., Park, J., Bennis, M., Kim, S.L.: XOR mixup: privacy-preserving data augmentation for one-shot federated learning. In: International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2020 (FL-ICML 2020) (2020)

30. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv. Neural Inf. Process. Syst. **33**, 596–608 (2020)
31. Tang, Z., Zhang, Y., Shi, S., He, X., Han, B., Chu, X.: Virtual homogeneity learning: defending against data heterogeneity in federated learning. arXiv preprint arXiv:2206.02465 (2022)
32. Wang, H., Muñoz-González, L., Eklund, D., Raza, S.: Non-IID data re-balancing at IoT edge with peer-to-peer federated learning for anomaly detection. In: Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks, pp. 153–163 (2021)
33. Wang, H., Kaplan, Z., Niu, D., Li, B.: Optimizing federated learning on non-IID data with reinforcement learning. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications, pp. 1698–1707. IEEE (2020)
34. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Adv. Neural Inf. Process. Syst. **33**, 7611–7623 (2020)
35. Wang, L., Lin, Z.Q., Wong, A.: COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. **10**(1), 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z
36. Wicaksana, J., Yan, Z., Yang, X., Liu, Y., Fan, L., Cheng, K.T.: Customized federated learning for multi-source decentralized medical image classification. IEEE J. Biomed. Health Inform. **26**, 5596–5607 (2022)
37. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
38. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739 (2018)
39. Yoon, T., Shin, S., Hwang, S.J., Yang, E.: FedMix: approximation of mixup under mean augmented federated learning. In: International Conference on Learning Representations (2020)
40. Yoshida, N., Nishio, T., Morikura, M., Yamamoto, K., Yonetani, R.: Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data. In: ICC 2020–2020 IEEE International Conference on Communications (ICC), pp. 1–7. IEEE (2020)
41. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: beyond empirical risk minimization. International Conference on Learning Representations (ICLR) (2018)
42. Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., Wu, Y.: FedDPGAN: federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia. Inf. Syst. Front. **23**, 1–13 (2021)
43. Zhang, W., Wang, X., Zhou, P., Wu, W., Zhang, X.: Client selection for federated learning with non-IID data in mobile edge computing. IEEE Access **9**, 24462–24474 (2021)
44. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-IID data. arXiv preprint arXiv:1806.00582 (2018)
45. Zhou, T., Zhang, J., Tsang, D.: FedFA: federated learning with feature anchors to align feature and classifier for heterogeneous data. arXiv preprint arXiv:2211.09299 (2022)
46. Zhu, H., Xu, J., Liu, S., Jin, Y.: Federated learning on non-IID data: a survey. Neurocomputing **465**, 371–390 (2021)
47. Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: International Conference on Machine Learning, pp. 12878–12889. PMLR (2021)