FEDERATED LEARNING WITH GAN-BASED DATA SYNTHESIS FOR NON-IID CLIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) has recently emerged as a popular privacy-preserving collaborative learning paradigm. However, it suffers from the non-IID (independent and identically distributed) data among clients. In this paper, we propose a novel framework, namely Synthetic Data Aided Federated Learning (SDA-FL), to resolve the non-IID issue by sharing differentially private synthetic data. Specifically, each client pretrains a local generative adversarial network (GAN) to generate synthetic data, which are uploaded to the parameter server (PS) to construct a global shared synthetic dataset. The PS is responsible for generating and updating high-quality labels for the global dataset via pseudo labeling with a confident threshold before each global aggregation. A combination of the local private dataset and labeled synthetic dataset leads to nearly identical data distributions among clients, which improves the consistency among local models and benefits the global aggregation. To ensure privacy, the local GANs are trained with differential privacy by adding artificial noise to the local model gradients before being uploaded to the PS. 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.

1 INTRODUCTION

The recent development of deep learning technologies has led to major breakthroughs in various domains. Meanwhile, the number of mobile clients (e.g., mobile phones, wearable devices, and smart sensors) is growing drastically. 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 by 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. Particularly, McMahan et al. (2017) introduced the Federated Averaging (FedAvg) algorithm that synchronously aggregates the updates of local models, while satisfying the basic requirements for privacy protection. Later, FedProx was proposed by Li et al. (2020) that adds a proximal term in the loss function to improve the statistical stability of the training process.

A key characteristic of federated learning is the non-IID (independent and identically distributed) data distribution among clients. Consequently, different clients learn from different data distributions, which leads to high heterogeneity among local models and degrades the effectiveness of model aggregation (McMahan et al., 2017). For extreme cases with highly skewed data distribution, traditional federated learning methods even lack a theoretical guarantee of convergence (Yu et al., 2020; Rothchild et al., 2020). Recent studies have attempted to address the non-IID issue by data sharing, which constructs a global shared dataset based on the local data to mitigate the data imbalance (Krizhevsky et al., 2009; Zhao et al., 2018; Yoshida et al., 2020; Oh et al., 2020). However, naive data sharing strategies directly share a percentage of local data among clients, which violates the privacy requirement in FL (Krizhevsky et al., 2009; Yoshida et al., 2020). Meanwhile, data augmentation approaches upload synthetic samples to the server without sharing private data, but these methods are still susceptible to data leakage without a privacy-preserving mechanism (Zhang et al.,

2021; Oh et al., 2020). Moreover, the limited local data samples severely affect the quality of synthetic data, and thus collaborative training cannot be performed just on these low-quality data. It is therefor important to investigate how to improve the quality of synthetic data via the FL framework, and how to effectively utilize the private dataset and synthetic dataset.

To alleviate the heterogeneity of the data distribution in FL without compromising privacy, we propose a novel framework, named Synthetic Data Aided Federated Learning (SDA-FL), which resolves the non-IID issue by sharing differentially private synthetic data. Under this framework, each client pretrains a local generative adversarial network (GAN) to generate synthetic data, thus avoiding sharing the raw data directly. A parameter server (PS) performs pseudo-labeling for the uploaded synthetic data and constructs a global shared dataset. In each global aggregation, the PS updates the pseudo labels of the synthetic dataset based on the received models and updates the global model with the confident synthetic data. This iterative labeling process constitutes a synthetic dataset with high-confidence labels, which is fed back to the clients to construct their training samples, including the local private data and synthetic data, follow an almost identical distribution. Meanwhile, to provide a privacy guarantee for the generative models, we adopt the widely used differential privacy (Dwork, 2008) to ensure that the actual training samples can not be revealed by adversaries. In particular, we use the Wasserstein GAN (WGAN) (Gulrajani et al., 2017) as the backbone and introduce the privacy budget by adding artificial noise to the gradients in the training process. The SDA-FL framework is mostly compatible with many existing FL methods and can be applied in both supervised and semi-supervised settings.

We evaluate the proposed framework on several standard datasets under both supervised learning and semi-supervised learning, which shows its effectiveness in resolving the non-IID distribution problem even in the extreme case (with only one class of real data at each client). We also investigate the tradeoff between the privacy budget and the FL performance, which demonstrates that this framework is fully functional under strict differential privacy protection requirement. Our major contributions are summarized as follows:

- We develop a global shared synthetic dataset by the differentially private GAN models. This data augmentation method effectively resolves the non-iid problem in FL and maintains the local data privacy.
- To utilize the global shared synthetic dataset efficiently, we propose a novel framework, SDA-FL, in which the PS provides and keeps updating high-confidence pseudo labels for the synthetic dataset iteratively, thus improving the performance of the global model.
- We evaluate our framework on several benchmark datasets under the high heterogeneous data distribution, where our framework consistently outperforms other baselines under both supervised and semi-supervised learning. Besides, our framework experiences a slight performance degradation with a stringent privacy budget.

2 Related Work

Non-IID Challenges in Federated Learning The pioneering work in FL is FedAvg (McMahan et al., 2017), which simply averages model parameters trained by different clients in an element-wise manner, weighted proportionately by the clients' data size. However, this algorithm often suffers from performance degradation when data is heterogeneous over local clients. Some studies proposed personalized layers to address this issue, which aim to adjust the local model according to the private dataset (Arivazhagan et al., 2019; Liang et al., 2020). One method called FedPer was introduced in Arivazhagan et al. (2019), which develops personalized predictors for classifications. Knowledge distillation (Deng et al., 2020; Lin et al., 2020) is also a promising idea of personalized federated learning, which transfers information from other clients to a certain client. Although the aforementioned methods outperform FedAvg on strongly non-iid data, the extra computational costs are not affordable by the resource-constrained IoT devices (Zhu et al., 2021). Besides, client clustering has also been proposed to construct a multi-center framework by grouping the clients into different clusters based on the model similarity (Ghosh et al., 2020; Kopparapu & Lin, 2020). As clients avoid the negative knowledge transfer from the dissimilar models (Pan & Yang, 2009), clustering methods can alleviate the negative effect of the non-iid data distribution. However, this method is not scalable as the communication cost increases rapidly with the number of clients.

Data Augmentation and Privacy Concern Recently, data sharing methods have received increasing attentions and achieved prominent performance in FL (Zhu et al., 2021; Zhao et al., 2018). Particularly, data augmentation approaches showed the potential capability to address the non-IID issue by modifying the local distributions without sharing private data (Duan et al., 2019; Zhang et al., 2018; Eaton-Rosen et al., 2020). A vanilla data augmentation method was proposed by Duan et al. (2019), in which each client performs local data augmentation to reach a balanced global distribution. Besides, several methods (Oh et al., 2020; Shin et al., 2020; Yoon et al., 2021) leveraged the Mixup technique (Zhang et al., 2018) to conduct data augmentation, where the clients share the blended local data and collaboratively construct a new global dataset to tackle the non-IID issue. However, frequent data exchange may be susceptible to privacy issues. Recently, the GAN-based data augmentation framework (Bowles et al., 2018; Yoshida et al., 2020) received much attention for reducing the degree of local data imbalance in FL. The general idea is to train a good generative model in the presence of non-IID data. Under this framework, each client sends its local seed samples to the server for the GAN training, and then a well-trained generator will be sent to all clients for model updating. However, directly sending local data samples to the server violates the data privacy requirement. Federated GAN models (Zhang et al., 2021) suggested an alternative that each client trains a local GAN based on the private data and uploads the parameters to the server. However, the white-box generators sent to the server are vulnerable to adversarial attacks (Chen et al., 2020).

3 PRELIMINARY

Federated Learning FedAvg (McMahan et al., 2017) is one of the representative training algorithms for in federated learning, and it trains a shared neural network without disclosing the raw data of clients. For every step $t = 0, \ldots, T-1$, every clients $k \in \mathbb{S}_t$ downloads the global model $\boldsymbol{w}_t$ and updates the local model with local dataset $\mathbb{D}_k$ via stochastic gradient descent (SGD), i.e., ${\boldsymbol{w}_{t+1}}^k \leftarrow {\boldsymbol{w}_t}^k - \eta_{t+1}\nabla\ell$. The updated local models are then sent back to the PS for a weighted aggregation $\boldsymbol{w}_{t+1} \leftarrow \frac{1}{K}\sum_{k\in\mathbb{S}_t} \boldsymbol{w}_{t+1}^k$. These procedures repeat until convergence or all the T training steps are exhausted.

Differentially Private Generative Adversarial network We design a differentially private generative adversarial network (DPGAN) for data augmentation and local data privacy protection. Specifically, Wasserstein GAN (WGAN) (Gulrajani et al., 2017) is adopted as the backbone of the generative model, since it resolves the problems of vanishing gradient and mode collapse. Besides, we add noise to the gradient in the generator training process, which provides privacy control in terms of differential privacy (Torkzadehmahani et al., 2019). The definition of differential privacy is as follows:

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

$$\mathbb{P}\left(\mathcal{A}_p(\mathcal{D}) \in \mathcal{S}\right) \leq e^{\epsilon} \cdot \mathbb{P}\left(\mathcal{A}_p\left(\mathcal{D}'\right) \in \mathcal{S}\right) + \delta. \tag{1}$$

Particularly, δ is an optimal probability that guarantees the validity of the inequality, 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)-differential privacy, we follow the work (Xie et al., 2018a) and add the Gaussian noise to the gradients updated in each discriminator training iteration. The relationship between the noise variance and differential privacy is shown as follows:

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

where q and n_d denote the sample probability and the number of discriminator training iterations in each inner loop, respectively.

Pseudo Labeling Although GAN-based data augmentation can generate synthetic data, the labels of the synthetic data are unknown, which hinders the training of the model through back-propagation. In semi-supervised learning, pseudo labeling is a widely used method in classification tasks to generate confident pseudo labels (Sohn et al., 2020). A hard pseudo label will be applied to

the unlabeled data if the maximum class probability (MCP) of the inference result (Corbière et al., 2019) exceeds a predefined threshold τ. In our experiments, the threshold is set to be 0.95.

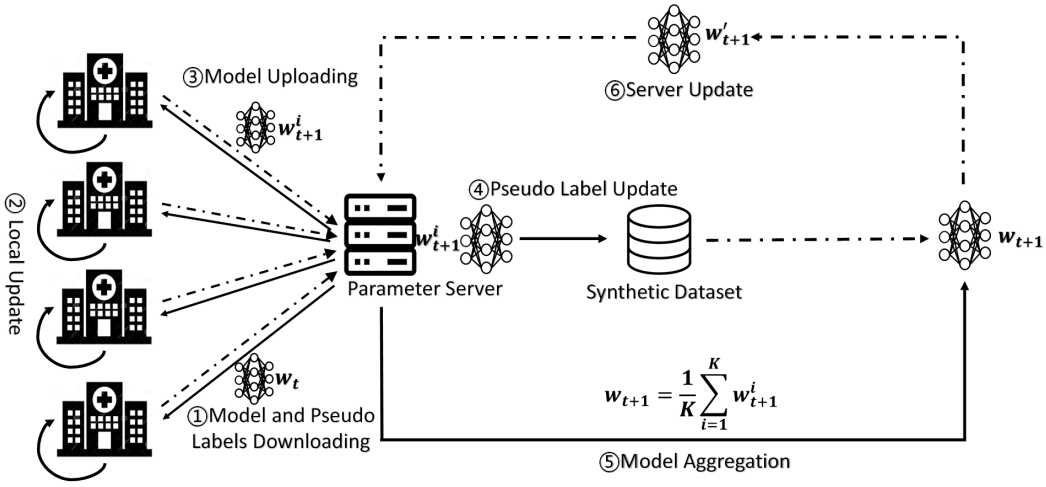


Figure 1: Overview of the proposed SDA-FL framework. Before the federated learning, the synthetic data from all clients are sent to the PS to construct a global synthetic dataset. In each iteration, every client first downloads the global model and updates the synthetic dataset for local training. The 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 system diagram is shown in Fig. 1, and key algorithmic innovations built upon the classic FL framework are elaborated below.

Global Synthetic Dataset Construction At the beginning of the federated training, each client pretrains a local GAN model to generate synthetic samples based on its private data. Then, the synthetic samples are sent to the PS to construct a global shared synthetic dataset. To effectively leverage the synthetic dataset for federated learning, we perform pseudo labeling for these samples, which is critical for the effectiveness of the proposed framework.

In federated learning, the local model is the most suitable network to predict the pseudo labels for the synthetic data generated by the corresponding local GAN, as they are both trained with the same data. In our framework, to improve the confidence level of the pseudo labels, we utilize the local models to perform the pseudo labeling for the unlabeled synthetic data. Specifically, after receiving the local model $\boldsymbol{w}_{t+1}^i$ in each aggregation step, the PS utilizes the model predictions to assign a pseudo label for each unlabeled synthetic instance if the maximum class probability is higher than a predefined threshold τ. In this way, we are gradually generating synthetic data samples with high-quality labels.

Synthetic Data Aided Model Training Augmented by the samples from the shared synthetic dataset that are with the confident labels, the data available for local training at different clients reach approximately homogeneous distribution. To make good use of the synthetic data, we leverage the $Mixup$ method (Zhang et al., 2018), which utilizes linear interpolation between the real sample (x_i, y_i) and the synthetic sample $(\hat{x}_i, \hat{y}_i)$ to augment the real data:

$$\begin{aligned} \bar{x} &= \lambda_1 x_i + (1-\lambda_1)\hat{x}_i, \\ \bar{y} &= \lambda_1 y_i + (1-\lambda_1)\hat{y}_i, \end{aligned} \tag{3}$$

Algorithm 1: Synthetic Data Augmented Federated Learning (SDA-FL)

Input: Local dataset $\mathbb{D}_k = \{\boldsymbol{X}_k, \boldsymbol{Y}_k\}$,
Pretrained generators $\mathbb{G} = \{\boldsymbol{G}_k\}$
$k \in \{1, \ldots, N\}$
Initialize w_0, Synthetic dataset $\hat{\mathbb{D}}_s = \{\hat{\boldsymbol{X}}_s, \hat{\boldsymbol{Y}}_s\}$ for server, $\hat{\mathbb{D}}_k = \{\hat{\boldsymbol{X}}_k, \hat{\boldsymbol{Y}}_k\}$ for clients
All clients upload the synthetic data generated by G_k to PS
for $t = 0, \ldots, T-1$ **do**
  $\mathbb{S}_t \leftarrow K$ clients are selected at random
  Send $\boldsymbol{w}_t$ and updated synthetic data to clients $k \in \mathbb{S}_t$
  Clients Side:
  for $k \in \mathbb{S}_t$ **do**
    Updata local synthesis dataset$(\hat{\boldsymbol{X}}_k, \hat{\boldsymbol{Y}}_k)$
    $\boldsymbol{w}_{t+1}^k \leftarrow$ *LocalUpdate*$(k, \boldsymbol{w}_t; \boldsymbol{X}_k, \boldsymbol{Y}_k, \hat{\boldsymbol{X}}_k, \hat{\boldsymbol{Y}}_k)$
  end
  Parameter Server Side:
  Pseudo Label Update
  $\boldsymbol{w}_{t+1} \leftarrow \frac{1}{K}\sum_{k\in\mathbb{S}_t} \boldsymbol{w}_{t+1}^k$
  $\boldsymbol{w}_{t+1} \leftarrow$ *ServerUpdate*$(\boldsymbol{w}_{t+1}; \hat{\boldsymbol{X}}_s, \hat{\boldsymbol{Y}}_s)$
end

Algorithm 2: Local Update

LocalUpdate$(k, \boldsymbol{w}_t; \boldsymbol{X}_k, \boldsymbol{Y}_k, \hat{\boldsymbol{X}}_k, \hat{\boldsymbol{Y}}_k)$ under SDA-FL (Algorithm 1):
$\boldsymbol{w} \leftarrow \boldsymbol{w}_t$
for $e = 0, \ldots, E-1$ **do**
  Split $\mathbb{D}_k$ into batches of size B
  for *batch*$(\boldsymbol{X}, \boldsymbol{Y})$ **do**
    Sample batch$(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}})$ from $(\hat{\boldsymbol{X}}_k, \hat{\boldsymbol{Y}}_k)$
    $\ell_1 = (1-\lambda_1)\ell\big(f((1-\lambda_1)\boldsymbol{X}; \boldsymbol{w}), \boldsymbol{Y}\big) + \lambda_1\ell\big(f(\lambda_1\hat{\boldsymbol{X}}; \boldsymbol{w}), \hat{\boldsymbol{Y}}\big)$
    $\ell_2 = \ell\big(f(\boldsymbol{X}; \boldsymbol{w}), \boldsymbol{Y}\big)$
    $\ell = \ell_1 + \lambda_2\ell_2$
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta_{t+1}\nabla\ell$
  end
end
return $\boldsymbol{w}$
The *ServerUpdate*$(\boldsymbol{w}_t; \hat{\boldsymbol{X}}_s, \hat{\boldsymbol{Y}}_s)$ procedure is the same as the *LocalUpdate* execpt the dataset

where λ_1 follows the Beta distribution $\big(\text{Beta}(\alpha, \alpha)\big)$ with $\alpha \in [0, 1]$. Combined with the cross-entropy loss, the Mixup loss for local update becomes:

$$\ell_1 = (1-\lambda_1)\ell\big(f((1-\lambda_1)\boldsymbol{X}_i; \boldsymbol{w}_i), \boldsymbol{Y}_i\big) + \lambda_1\ell\big(f(\lambda_1\hat{\boldsymbol{X}}_i; \boldsymbol{w}_i), \hat{\boldsymbol{Y}}_i\big). \tag{4}$$

In addition, to keep more information of the local clients, another cross-entropy loss term is introduced for the real data as follows:

$$\ell_2 = \ell\big(f(\boldsymbol{X}_i; \boldsymbol{w}_i), \boldsymbol{Y}_i\big). \tag{5}$$

Then SGD is applied to update the local model:

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i - \eta_{t+1}\nabla(\ell_1 + \lambda_2\ell_2), \tag{6}$$

where λ_2 is a hyperparameter to control the retention of the local information and will be set in different values in experiments. More details about the local update are shown in Algorithm 2.

In contrast to the traditional federated learning where the PS does not have access to any data to update the global model, the PS in our framework can keep the entire global synthetic dataset $\hat{\mathbb{D}}_s$ and uses it to update the global model, thereby helping to enhance the performance.

Interplay between Model Training and Synthetic Dataset Updating Because of the non-iid problem, there is a distinct difference in the weights of the local models among clients in federated learning. Furthermore, simply sharing the synthetic data with the PS and other clients will not result in the good training performance of the global model due to the lack of confident pseudo labels. Our framework, SDA-FL, deeply integrates federated learning and synthetic data, which effectively alleviates the non-IID issue and helps to train a good global model. As shown in Fig. 1, after creating the global synthetic dataset, it will be shared with all the clients so that the local data, including both the real and the synthetic data, approach an IID distribution. Without a doubt, the supplement of synthetic data results in the improvement of the local models. After being uploaded to the PS, the updated local models are used to perform the pseudo labeling and update

Table 1: Test accuracy of federated supervised learning on MNIST.

#class/client	FedAvg	Fedprox	Scaffold	Ours
1	83.44%	84.17%	25.39%	**98.19**%
2	97.61%	97.55%	94.17%	**98.26**%
3	98.42%	98.38%	96.89%	**98.50**%

Table 2: Test accuracy of federated supervised learning on FashionMNIST.

#class/client	FedAvg	Fedprox	Scaffold	Ours
1	16.50%	57.14%	56.80%	**85.70**%
2	73.50%	75.76%	70.82%	**86.55**%
3	82.47%	83.43%	77.68%	**87.06**%

the global synthetic dataset by the PS. The stronger local models can boost the confidence of the pseudo label. With the improvement of labeling for the global synthetic dataset, the PS can utilize it to update the global model and all the clients can improve their local models subsequently at the next-round training. Thus there is an interplay between the model update and the synthetic dataset at every training round: global synthetic dataset $\rightarrow$ global model update $\rightarrow$ local model updaet $\rightarrow$ labeling for global synthetic dataset, which results in a converged and well-perform model.

SDA-FL vs. Traditional FL From the above descriptions, the proposed SDA-FL framework introduces two additional operations at the server side, namely, pseudo label update of the shared synthetic dataset and server-side model update, as shown in Fig. 1. These two main innovations contribute to the performance improvement with non-IID data. In traditional FL algorithms (Li et al., 2020; Karimireddy et al., 2020), clients update their models based on local private data, which may lead to performance degradation when data are heterogeneous over clients. In our framework, the local dataset is augmented by the GAN-based synthetic samples, which alleviates the non-IID problem. Furthermore, the PS in the traditional FL system only performs model aggregation based the uploaded local models. In contrast, our framework generates high-confidence synthetic data to update and improve the global model at the PS. Overall, with a shared synthetic dataset and an effective mechanism for pseudo label update, SDA-FL is able to overcome the heterogeneous data distributions among clients and enhance the model update at the server. We envision that this framework can be applied to develop other data augmentation-based methods for more efficient federated learning.

5 Experiments

In this section, we evaluate the proposed SDA-FL framework in the presence of non-IID data for both supervised learning and semi-supervised learning. The experimental results on various benchmark datasets demonstrate the effectiveness of the proposed framework compared with baseline methods. Besides, an ablation study is conducted to illustrate the impact of the privacy budget on the generated samples and the model convergence.

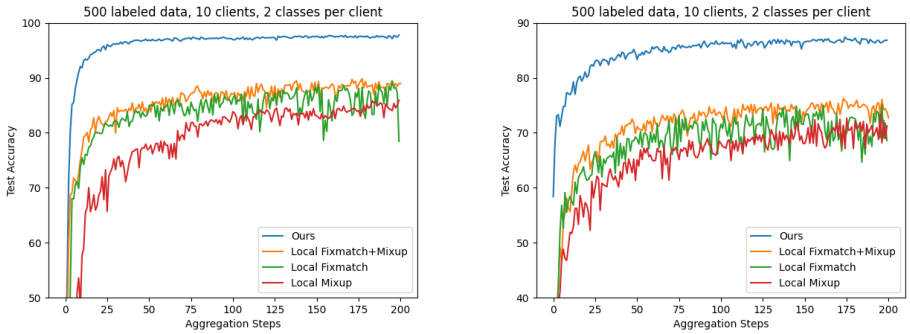
5.1 Experimental Setup

Datasets We implement the SDA-FL framework in the scenario where all the real data are stored in clients and the PS only has access to the synthetic data. We use three benchmark datasets, including MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), and Cifar-10 (Krizhevsky, 2009) to test the proposed method. There are 60,000 training samples and 10,000 testing samples in MNIST dataset, while both FashionMNIST and Cifar-10 datasets are divided into 50,000 training images and 10,000 testing images.

Models For fair comparisons, we adopt a simple CNN model that consists of two convolution layers and two fully-connected layers for the MNIST and FashionMNIST classification tasks. Meanwhile, ResNet18 (He et al., 2016) is used for classifying the each Cifar-10 image dataset.

Table 3: Test accuracy of federated supervised learning on Cifar-10.

Algorithms	FedAvg	Fedprox	Scaffold	Ours
Accuracy	61.28%	63.16%	60.78%	**67.51**%



(a) MNIST (b) FashionMNIST

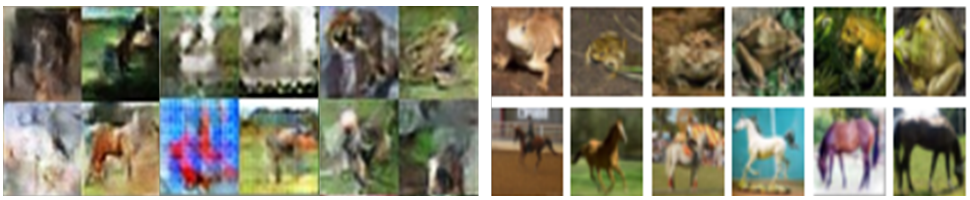
Figure 2: Test accuracy of different methods for federated semi-supervised learning on MNIST and FashionMNIST.

Tasks **1) Federated supervised learning:** We compare our framework with FedAvg (McMahan et al., 2017), Fedprox (Li et al., 2020), and Scaffold (Karimireddy et al., 2020) on MNIST, FashionMNIST, and Cifar-10 datasets. To configure the non-iid setting, given the number of classes per client as C and with K clients, the whole dataset will be split into $K * C$ subsets, each with a single class. Then all the subsets of data will be shuffled and distributed to different clients. **2) Federated semi-supervised learning:** We extend our framework to the semi-supervised learning setting by performing pseudo labeling for the unlabeled data. In this experiment, every client only keeps 2 classes of data, containing 50 labeled instances, and the rest are unlabeled. To show the effectiveness of our SDA-FL method, we compare with Local Fixmatch+Mixup (Diao et al., 2021), Local Fixmatch (Sohn et al., 2020), and Local Mixup (Zhang et al., 2018) (mixup between labeled data and unlabeled data). **3) The ablation study of privacy budget:** To investigate the influence of the differential privacy on model convergence, we also validate our framework under different privacy budgets. The experimental setting is the same as the supervised learning experiment with 2 classes of data at every client. **4) The effectiveness of the Synthetic data:** To investigate the effectiveness of the synthetic data, we compare our framework with NaiveMix and Fedmix (Yoon et al., 2021), both of which upload the average of M_k real images of each client to resolve the non-IID problem.

5.2 Performance of SDA-FL on Non-IID Federated Learning

Federated Supervised Learning We compare the performance of different methods under the same federated settings. With a varying number of classes per client, the experimental results are shown in Table 1 and Table 2, and our framework outperforms other methods by a large margin. Particularly, the results show a remarkable improvement in test accuracy when the degree of non-iidness is high, which is attributed to the GAN-based data augmentation that mitigates the detrimental effects of the data heterogeneity. The experiment results on Cifar-10 also show that our framework outperforms the classic Fedprox algorithm by 4.4% as shown in Table 3.

Federated Semi-Supervised Learning To show the robustness and generalizability of our framework, we extend the experiments to federated semi-supervised learning and compare with three baselines. The results in Fig. 2 show that our framework achieves faster convergence and better performance than other algorithms. Particularly, compared with Local Fixmatch+Local Mixup, our method improves the accuracy by 5% in the FashionMNIST classification. This is because the proposed pseudo labeling mechanism can provide high-quality labels for the synthetic dataset and unlabeled dataset, which benefits the federated training process.



(a) Synthetic data (b) Original Cifar-10 dataset

Figure 3: Visual comparison between synthetic data and the original Cifar-10 dataset.

Table 4: Test accuracy of federated semi-supervised learning on the Cifar-10 dataset

Algorithms	Ours	Fixmatch+Mixup	Fixmatch	Mixup
Accuracy	**51.68**%	25.52%	29.5%	18.08%

The experiment results on the Cifar-10 dataset are shown in Table 4. The baseline algorithms are not able to train a usable global model with less than 30% of test accuracy, while our framework converges in this challenging scenario and improves the test accuracy by a significant margin.

Effect of privacy budget The privacy budget ϵ provides the privacy protection for the local data during data synthesis, but a small ϵ affects the quality of the synthetic data and may degrade the training performance. To investigate the impact of the privacy budget, we conduct experiments with different values of ϵ on the MNIST, FashionMNIST, and Cifar-10 datasets. As shown in Fig. 4, the proposed framework suffers from performance degradation with a small privacy budget. Compared with a protection-free scenario ($\epsilon = \infty$), there are around 4% and 10% accuracy drop on MNIST and FashionMNIST datasets, respectively, when setting a strict privacy budget ($\epsilon = 3$) (Xie et al., 2018b). However, our framework still enjoys superior learning performance compared to other methods as shown in Table 3.

For the Cifar-10 classification task, we observe that the generators are sensitive to the Gaussian noise added to the gradient, and thus a large privacy budget is adopted. Tabel 5 shows that our framework outperforms Fedmix with the same privacy budget. To illustrate the privacy guarantee on the generators in our framework, we show some synthetic samples in Fig. 3(a), which are extracted from the generators trained with $\epsilon = 200$. In comparison with the samples from the private datasets, the synthetic images are blurry and noisy, which means the GAN models are not memorizing the real data, and the privacy is not undermined. Meanwhile, the results in Table 5 show that our framework outperforms other algorithms in terms of learning accuracy, which indicates that the synthetic images with low perceptual quality are able to capture the necessary information for the model training.

Effect of the synthetic data To prove that the synthetic data with privacy protection outperforms the information sharing method via averaging multiple samples (Yoon et al., 2021), we set $\lambda_1 = 0.01$ and $\lambda_2 = 0$ following the Fedmix algorithm setting, and choose $M_k = 9$. The performance comparison between our framework and Fedmix is shown in Table 6. Under the same experimental setting except for the augmentation data, our framework outperforms NaiveMix, which indicates that augmentation via synthetic data is more effective than averaging real data samples. Although Fedmix adds another derivative term when averaging real images to explicitly handle the non-iid problem, our framework still achieves better performance.

Table 5: Test accuracy on Cifar-10 under various privacy budgets

#Privacy Budget	$\epsilon = 50$	$\epsilon = 100$	$\epsilon = 200$	$\epsilon = \infty$
Accuracy	65.87%	66.41%	67.23%	67.51%

Table 6: Test accuracy comparison of NaiveMix and Fedmix on Cifar-10

Algorithms	NaiveMix	Fedmix	Ours ($\epsilon = 200$)
Accuracy	64.39%	65.76%	**67.23**%

(a) MNIST (b) FashionMNIST

Figure 4: Test accuracy on MNIST and FashionMNIST under various privacy budget

6 Conclusion

We proposed a new data augmentation method to resolve the imbalanced data distribution problem in federated learning, by introducing pretrained GANs to construct a differentially private global shared dataset. To effectively utilize the synthetic data, a novel framework, named Synthetic Data Aided Federated Learning (SDA-FL), was developed, which generates and updates the confident pseudo labels for the synthetic data samples. By supplementing the non-IID private datasets with high-quality synthetic data samples, this framework effectively improves the consistency among local models. Experiment results showed that SDA-FL outperforms many existing baselines by remarkable margins in various image classification tasks and is still functional under strict differential privacy protection.

References

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 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.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Conference on Neural Information Processing Systems (NIPS)*, pp. 2902–2913, 2019.

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

Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Communication efficient semi-supervised federated learning with unlabeled clients. *arXiv e-prints*, pp. arXiv–2106, 2021.

Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep

learning applications. In *2019 IEEE 37th international conference on computer design (ICCD)*, pp. 246–254. IEEE, 2019.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Z. Eaton-Rosen, Felix J. S. Bragman, Sébastien Ourselin, and M. Cardoso. Improving data augmentation for medical image segmentation. *Conference on Medical Imaging with Deep Learning (MIDL)*, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

Kavya Kopparapu and Eric Lin. Fedfmc: Sequential efficient federated learning on non-iid data. *arXiv preprint arXiv:2006.10937*, 2020.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2020.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Seungeun Oh, Jihong Park, Eunjeong Jeong, Hyesung Kim, Mehdi Bennis, and Seong-Lyun Kim. Mix2fld: Downlink federated learning after uplink federated distillation with two-way mixup. *IEEE Communication Letters*, 2020.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pp. 8253–8265. PMLR, 2020.

MyungJae Shin, Chihoon Hwang, Joongheon Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Xor mixup: Privacy-preserving data augmentation for one-shot federated learning. *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2020 (FL-ICML'20)*, 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

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

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv e-prints*, pp. arXiv–1802, 2018a.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018b.

Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv preprint arXiv:2107.00233*, 2021.

Naoya Yoshida, Takayuki Nishio, Masahiro Morikura, Koji Yamamoto, and Ryo Yonetani. 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.

Felix Yu, Ankit Singh Rawat, Aditya Menon, and Sanjiv Kumar. Federated learning with only positive labels. In *International Conference on Machine Learning*, pp. 10946–10956. PMLR, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations (ICLR)*, 2018.

Longling Zhang, Bochen Shen, Ahmed Barnawi, Shan Xi, Neeraj Kumar, and Yi Wu. Feddpgan: Federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. *Information Systems Frontiers*, pp. 1–13, 2021.

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.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *arXiv preprint arXiv:2106.06843*, 2021.