Delving into the Adversarial Robustness of Federated Learning

Jie Zhang[1*] **Bo Li**[2*‡] **Chen Chen**[3] **Lingjuan Lyu**[3‡]
Shuang Wu[2] **Shouhong Ding**[2] **Chao Wu**[1‡]
[1]Zhejiang University [2]Youtu Lab, Tencent [3]Sony AI
{zj_zhangjie, chao.wu}@zju.edu.cn
{libraboli, calvinwu, ericshding}@tencent.com, {chen.chen, Lingjuan.Lv}@sony.com

Abstract

In Federated Learning (FL), models are as fragile as centrally trained models against adversarial examples. However, the adversarial robustness of federated learning remains largely unexplored. This paper casts light on the challenge of adversarial robustness of federated learning. To facilitate a better understanding of the adversarial vulnerability of the existing FL methods, we conduct comprehensive robustness evaluations on various attacks and adversarial training methods. Moreover, we reveal the negative impacts induced by directly adopting adversarial training in FL, which seriously hurts the test accuracy, especially in non-IID settings. In this work, we propose a novel algorithm called Decision Boundary based Federated Adversarial Training (DBFAT), which consists of two components (local re-weighting and global regularization) to improve both **accuracy** and **robustness** of FL systems. Extensive experiments on multiple datasets demonstrate that DBFAT consistently outperforms other baselines under both IID and non-IID settings.

Introduction

Nowadays, end devices are generating massive amounts of potentially sensitive user data, raising practical concerns over security and privacy. Federated Learning (FL) (McMahan et al. 2017) emerges as a privacy-aware learning paradigm that allows multiple clients to collaboratively train neural networks without revealing their raw data. Recently, FL has attracted increasing attention from different areas, including medical image analysis (Liu et al. 2021a; Chen et al. 2021b), recommender systems (Liang, Pan, and Ming 2021; Liu et al. 2021b), natural language processing (Zhu et al. 2020; Wang et al. 2021), etc.

Prior studies have demonstrated that neural networks are vulnerable to evasion attacks by adversarial examples (Goodfellow, Shlens, and Szegedy 2014) during inference time. The goal of inference-time adversarial attack (Li et al. 2021a; Chen et al. 2022c; Zhang et al. 2022b; Chen et al. 2022b) is to damage the global model by adding a carefully generated imperceptible perturbation on the test examples. As shown in Table 1, federated models are as fragile to adversarial examples as centrally trained models (i.e. zero accuracy under PGD-40 attack (Madry et al. 2017)). Hence, it is also important to consider how to defend against adversarial attacks in federated learning.

There are several works that aim to deal with adversarial attacks in FL (Zhang et al. 2022c,a), i.e, federated adversarial training (FAT) (Zizzo et al. 2020; Hong et al. 2021; Shah et al. 2021; Chen, Zhang, and Lyu 2022; Chen et al. 2022a). (Zizzo et al. 2020) and (Hong et al. 2021) proposed to conduct adversarial training (AT) on a proportion of clients but conduct plain training on other clients. (Shah et al. 2021) investigated the impact of local training rounds in FAT. Nevertheless, these methods all ignore the issue that the clean accuracy of federated adversarial training is very low.

To further show the problems of federated adversarial training, we first begin with the comparison between the plainly-trained models and AT-trained (Madry et al. 2017) models in both the IID (Independent and Identically Distributed) and non-IID FL settings, measured by clean accuracy A_{cln} and robust accuracy A_{rob}, respectively. We show the test accuracy of plain training and adversarial training (AT) on CIFAR10 dataset under both IID and non-IID FL settings in Fig. 1 (left sub-figure). We summarize some valuable observations as follows: 1) Compared with the plainly-trained models, AT-trained models achieve a lower accuracy, which indicates that directly adopting adversarial training in FL can hurt A_{cln}; 2) A_{cln} drops heavily for both the plainly-trained models and AT-trained models under non-IID distribution, which is exactly the challenge that typical federated learning with heterogeneous data encountered (Zhao et al. 2018); 3) The performance of AT-trained models with non-IID data distribution decrease significantly compared with IID data distribution. Motivated by these observations, we focus on improving both adversarial robustness and clean accuracy of adversarial training in FL, i.e., we aim to increase A_{cln} while keeping A_{rob} as high as possible.

To achieve this goal, in this paper, we investigate the impact of decision boundary, which can greatly influence the performance of the model in FAT. Specifically, 1) we apply adversarial training with a re-weighting strategy in local update to get a better A_{rob}. Our method takes the limited data of each client into account, those samples that are close to/far from the decision boundary are assigned larger/smaller weight. 2) Moreover, since the global model in FL has a

[*] Equal contribution. Work done during Jie Zhang's internship at Tencent Youtu Lab and partly done at Sony AI.

[‡] Corresponding author.

Table 1: The accuracy (%) is tested under PGD-40 attack (Madry et al. 2017). For MNIST, FMNIST, CIFAR10, ImageNet-12, CIFAR100, and Tiny-ImageNet, the perturbation bound is $\{0.3, 32/255, 0.031, 0.031, 0.031, 0.031\}$, respectively. A_{cln} and A_{rob} refer to clean accuracy and robust accuracy.

Type	Dataset	MNIST	FMNIST	ImageNet-12	CIFAR10	CIFAR100	Tiny-ImageNet
Centralized	A_{cln}	99.42	92.47	78.96	94.26	86.93	57.93
	A_{rob}	0	0	0	0	0	0
Federated	A_{cln}	99.01	88.51	71.65	85.81	81.28	49.79
	A_{rob}	0	0	0	0	0	0

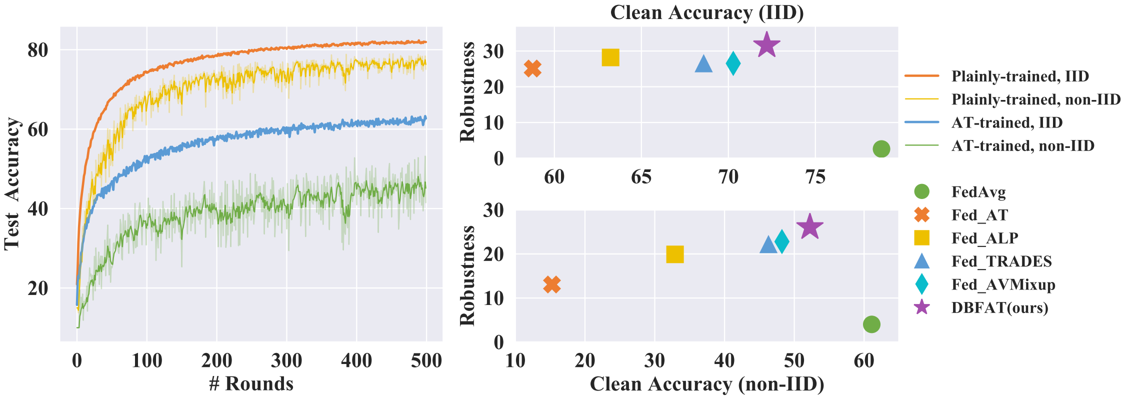

Figure 1: **Left:** Test accuracy reduces for plainly trained model and adversarially trained model under non-IID data. Meanwhile, adversarial training hurts the performance. **Right:** Evaluations on CIFAR10 for both accuracy and robustness, including several state-of-the-art defense methods combined with FL. Our method outperforms existing baselines on both metric dimensions.

more accurate decision boundary through model aggregation, we take advantage of the logits from the global model and introduce a new regularization term to increase A_{cln}. This regularization term aims to alleviate the accuracy reduction across distributed clients.

We conclude our major contributions as follows:

- We conduct systematic studies on the adversarial robustness of FL, and provide valuable observations from extensive experiments.
- We reveal the negative impacts of adopting adversarial training in FL, and then propose an effective algorithm called Decision Boundary based Federated Adversarial Training (DBFAT), which utilized local re-weighting and global regularization to improve both the accuracy and robustness of FL systems.
- Extensive experiments on multiple datasets demonstrate that our proposed DBFAT consistently outperforms other baselines under both IID and non-IID settings. We present the performance of our method in Fig. 1 (right sub-figure), which indicates the improvement in both robustness and accuracy of adversarial training in FL.

Related Works

Federated Learning. Following the success of DNNs in various tasks (Li et al. 2019; Li, Sun, and Guo 2019; **?**; Huang et al. 2022b,a; Dong et al. 2021), FL has attracted increasing attention. A recent survey has pointed out that existing FL systems are vulnerable to various attacks that aim to either compromise data privacy or system robustness (Lyu et al. 2022). In particular, robustness attacks can be broadly classified into training-time attacks (data poisoning and model poisoning) and inference-time attacks (evasion attacks, i.e., using adversarial examples to attack the global model during inference phase). In FL, the architectural design, distributed nature, and data constraints can bring new threats and failures (Kairouz 2021).

Adversarial Attacks. The white-box attacks have access to the whole details of threat models, including parameters and architectures. Goodfellow et al. (Goodfellow, Shlens, and Szegedy 2014) introduced the Fast Gradient Sign Method (FGSM) to generate adversarial examples, which uses a single-step first-order approximation to perform gradient ascent. Kurakin et al. (Kurakin, Goodfellow, and Bengio 2017) iteratively applied FGSM with a small step-size to develop a significantly stronger multi-step variant, called Iterative FGSM (I-FGSM). Based on these findings, more powerful attacks have been proposed in recent years including MIM (Dong et al. 2018), PGD (Madry et al. 2017), CW (Carlini and Wagner 2017), and AA (Croce and Hein 2020).

Adversarial Training. Adversarial training has been one of the most effective defense strategies against adversarial attacks. Madry et al. (Madry et al. 2017) regarded adversarial training as a min-max formulation using empirical risk minimization under PGD attack. Kannan et al. (Kannan, Kurakin, and Goodfellow 2018) presented adversarial logit pairing (ALP), a method that encourages logits for pairs of

Table 2: An empirical study on the adversarial robustness of FL, measured by various combination of defense methods and FL algorithms. We report the clean accuracy and robust accuracy, respectively. Best results are in bold.

Type	IID								Non-IID							
Methods	FedAvg		FedProx		FedNova		Scaffold		FedAvg		FedProx		FedNova		Scaffold	
Performance	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}	A_{cln}	A_{rob}
PGD-AT	57.99	31.95	58.17	32.06	58.45	31.74	56.84	29.26	46.84	26.79	48.03	27.46	46.95	26.54	42.44	27.19
ALP	62.81	31.84	62.88	31.20	62.91	31.79	60.30	29.58	56.16	**28.78**	55.79	**29.06**	55.80	**29.18**	48.29	26.56
TRADES	64.94	**32.93**	64.29	32.97	64.46	33.29	63.14	**33.58**	60.94	27.06	61.05	27.94	60.34	28.78	59.53	27.78
MMA	65.14	30.29	63.65	31.29	**65.27**	29.31	64.28	32.98	59.69	28.64	60.17	28.09	61.03	28.47	61.53	28.13
AVMixup	**66.14**	32.27	**65.12**	**33.19**	65.14	**33.75**	**65.11**	33.24	**61.17**	28.56	**61.47**	28.34	**62.04**	28.12	**61.91**	**28.81**

examples to be similar, to improve robust accuracy. To quantify the trade-off between accuracy and robustness, Zhang et al. (Zhang et al. 2019) introduced a TRADES loss to achieve a tight upper bound on the gap between clean and robust error. Based on the margin theory and soft-labeled data augmentation, Ding et al. (Ding et al. 2020) proposed Max-Margin Adversarial (MMA) training and Lee et al. (Lee, Lee, and Yoon 2020) introduced Adversarial Vertex mixup (AVmixup).

Federated Adversarial Training. In terms of the adversarial robustness, Zizzo et al. (Zizzo et al. 2020) investigated the effectiveness of the federated adversarial training protocol for idealized federated settings, and showed the performance of their models in a traditional centralized setting and a distributed FL scenario. Zhou et al. (Zhou et al. 2022) decomposed the aggregation error of the central server into bias and variance. However, all these methods sacrificed clean accuracy (compared to plainly trained models) to gain robustness. In addition, certified defense (Chen et al. 2021a) against adversarial examples in FL is another interesting direction, which will be discussed in the future.

Adversarial Robustness of FL

In this section, we briefly define the goal of federated adversarial training. Then we conduct a systematic study on some popular federated learning algorithms with the combination of various adversarial training methods and evaluate their robustness under several attacks. Besides, we further reveal the challenges of adversarial training in non-IID FL.

Problem Definition

In typical federated learning, training data are distributed across all the K clients, and there is a central server managing model aggregations and communications with clients. In general, federated learning attempts to minimize the following optimization:

$$\min_{w} f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w). \quad (1)$$

Here, we denote that the global approximate optimal is a sum of local objectives weighted by the local data size n_k, and n is the total data size of all clients that participate in a communication round. Moreover, each local objective measures the empirical risk over possibly different data distributions D_k, which can be expressed as:

$$F_k(w) := \mathbb{E}_{x_k \sim \mathcal{D}_k} \left[f_k \left(w; x_k \right) \right]. \quad (2)$$

Let x denote the original image, x^{adv} denote the corresponding adversarial example, and δ denote the perturbation added on the original image, then $x^{adv} = x + \delta$. To generate powerful adversarial examples, we attempt to maximize the loss $L(x + \delta; w)$, where L is the loss function for local update.

To improve the robustness of the neural networks, many adversarial defense methods have been proposed. Among them, adversarial training (Carlini and Wagner 2017) is one of the most prevailing and effective algorithms. Combined with adversarial training, the local objective becomes solving the following min-max optimization problem:

$$F_k(w) = \min \mathbb{E}_{x_k \sim \mathcal{D}_k} \left[\max_{\|x^{adv} - x\|_\infty \leq \delta} L(w, x^{adv}, y) \right]. \quad (3)$$

The inner maximization problem aims to find effective adversarial examples that achieve a high loss, while the outer optimization updates local models to minimize training loss.

In this work, we conduct a systematic study on several state-of-the-art FL algorithms including FedAvg (McMahan et al. 2017), FedProx (Li et al. 2018), FedNova (Wang et al. 2020) and Scaffold (Karimireddy et al. 2020), and explore their combinations with AT methods to defend against adversarial attacks. We report detailed results in Table 2, here robustness is averaged over four popular attacks (FGSM (Kurakin, Goodfellow, and Bengio 2017), MIM (Dong et al. 2018), PGD (Madry et al. 2017), and CW (Carlini and Wagner 2017)). Besides, we implement some prevailing adversarial training methods including PGD_AT (Madry et al. 2017) , TRADES (Zhang et al. 2019), ALP (Kannan, Kurakin, and Goodfellow 2018), MMA (Ding et al. 2020) and AVMixup (Lee, Lee, and Yoon 2020). We observe that there is no federated adversarial learning algorithm that can outperform all the others in all cases. Moreover, the clean accuracy drops heavily under non-IID distribution. As such, we are motivated to develop a more effective method. Due to the similar performance of these FL methods observed from Table 2, we design our method based on FedAvg – a representative algorithm in FL.

Adversarial Traning with non-IID Data

Federated learning faces the statistical challenge in real-world scenarios. The IID data makes the stochastic gradient as an unbiased estimate of the full gradient (McMahan et al. 2017). However, the clients are typically highly heterogeneous with various kinds of non-IID settings, such as

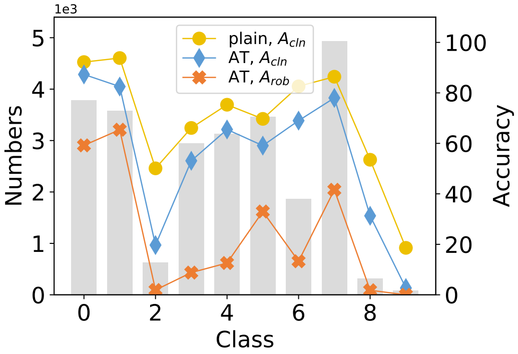

Figure 2: Test accuracy on a randomly selected client.

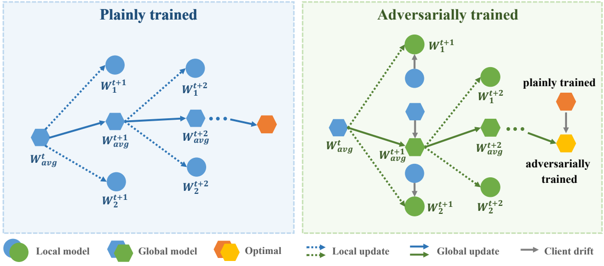

Figure 3: Plain training and adversarial training under non-IID setting. Compared with plainly trained situation, the aggregation of adversarially trained models can lead to a more biased model which enlarges accuracy gap. Consequently, it results in poor consistency between different clients.

label skewness and feature skewness (Li et al. 2021b). According to previous studies (Wang et al. 2020; Karimireddy et al. 2020), the non-IID data settings can degrade the effectiveness of the deployed model.

Similarly, due to the non-IID data, the performance of AT may vary widely across clients. To better understand the challenge of adversarial training with non-IID data, we examine the performance of both clean accuracy and robustness on a randomly selected client and report the results in Fig. 2. Observed from Fig. 2, we can find that: 1) A_{cln} on the plainly trained model drops from majority classes to minority classes, which is exactly what traditional imbalanced learning attempts to solve; 2) A similar decreasing tendency reasonably occurs in A_{rob}. It is obvious that adopting adversarial training in federated learning with non-IID data is more challenging.

According to above observations, we conjecture that AT-trained local models with imbalanced data lead to a more biased decision boundary than plainly trained ones. Since adversarial examples need a larger number of epochs to achieve near-zero error (Zhang et al. 2021), it becomes harder to fit adversarial examples than clean data. However, for the local client itself, imbalanced clean data generates imbalanced adversarial examples, making it more difficult for training and enlarging the accuracy gap, which can reduce the performance both in accuracy and robustness. In Fig. 3, we also show the differences between plain training and adversarial training in federated settings. Compared with the plainly trained models, the aggregation of adversarially trained models can enlarge the accuracy gap, which results in poor consistency between different clients. To overcome this problem, we propose a novel method to utilize local re-weighting and global regularization to improve both the accuracy and robustness of FL systems.

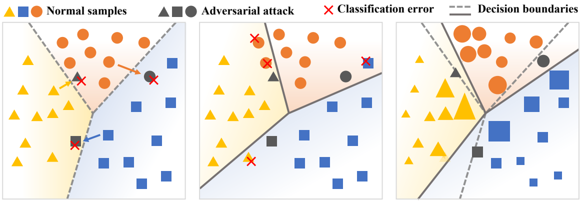

Figure 4: **Left panel:** Decision boundary of plainly trained model. **Middle panel:** Decision boundary of AT-trained model. **Right panel:** Decision boundary of DBFAT-trained model. We use the dotted line to represent the boundary of the clean model, and solid line to represent the boundary of the robust model. The size of the shape represents the value of the weight. Those samples that are close to/far from boundary are assigned larger/smaller weight. The decision boundary of DBFAT-trained model (see the right sub-figure) can achieve a higher A_{rob} and meanwhile maintain A_{cln}.

Methodology

The generalization performance of a neural network is closely related to its decision boundary. However, models trained in the federated setting are biased compared with the centrally trained models. This is mainly caused by heterogeneous data and objective inconsistency between clients (Kairouz 2021). Moreover, a highly skewed data distribution can lead to an extremely biased boundary (Wang et al. 2020). We tackle this problem in two ways: 1) locally, we take full advantage of the limited data on the distributed client; 2) globally, we utilize the information obtained from the global model to alleviate the biases between clients.

Subsequently, we propose a simple yet effective approach called Decision Boundary based Federated Adversarial Training (DBFAT), which consists of two components. For local training, we re-weight adversarial examples to improve robustness; while for global aggregation, we utilize the global model to regularize the accuracy for a lower boundary error A_{bdy}. We show the training process of DBFAT in the supplementary and illustrate an example of the decision boundary of our approach in Fig. 4.

Re-weighting with Limited Data

Adversarial examples have the ability to approximately measure the distances from original inputs to a classifier's decision boundary (Heo et al. 2018), which can be calculated by the least number of steps that iterative attack (e.g. PGD attack (Madry et al. 2017)) needs in order to find its misclassified adversarial variant. To better utilize limited adversarial examples, we attempt to re-weight the adversarial examples to guide adversarial training. For clean examples that

Table 3: Loss functions of different adversarial training methods.

Defense	Loss Function
PGD_AT	$\mathrm{CE}\left(f\left(x^{adv}\right), y\right)$
ALP	$\mathrm{CE}\left(\mathrm{f}\left(\mathrm{x}^{\mathrm{adv}}\right), \mathrm{y}\right) + \beta \cdot \left\Vert f\left(x^{adv}\right) - f(x)\right\Vert_2^2$
TRADES	$\mathrm{CE}\left(\mathrm{f}(\mathrm{x}), \mathrm{y}\right) + \beta \cdot \mathrm{KL}\left(f\left(x^{adv}\right) \Vert f(x)\right)$
MMA	$\mathrm{CE}\left(\mathrm{f}\left(\mathrm{x}^{\mathrm{adv}}\right), \mathrm{y}\right) \cdot \mathbb{R}\left(\mathrm{h}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{y}\right) + \mathrm{CE}\left(\mathrm{f}(\mathrm{x}), \mathrm{y}\right) \cdot \mathbb{R}\left(\mathrm{h}_{\boldsymbol{\theta}}(\mathbf{x}) \neq \mathrm{y}\right)$
AVMixup	$\mathrm{CE}\left(f\left(x^{av}\right), y^{av}\right)$
DBFAT(ours)	$\rho \cdot \mathrm{CE}(f(x^{adv}), y) + \beta \cdot \mathbf{KL}\left(f\left(x^{adv}\right) \Vert f^{glo}(x)\right)$

are close to the decision boundary, we assign larger weights; while those examples that are far from the boundary are assigned with smaller weights.

In this paper, we use PGD-S to approximately measure the geometric distance to the decision boundary, S denotes the number of maximum iteration. We generate adversarial examples as follows (Madry et al. 2017):

$$x^{adv} \leftarrow \Pi_{\mathcal{B}[x,\epsilon]}\left(x^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{x^{adv}} \ell(x^{adv}, y))\right). \quad (4)$$

Here $\Pi_{\mathcal{B}[x,\epsilon]}$ is the projection function that projects the adversarial data back into the ϵ-ball centered at natural data, α is the steps size, ϵ is perturbation bound.

We find the minimum step d, such that after d step of PGD, the adversarial variant can be misclassified by the network, i.e., $arg\ max_c f^{(c)}(x^{adv}) \neq y$, where $f^{(c)}(x^{adv})$ is the logits of the c-th label.

In this way, given a mini-batch samples $\{(x_i, y_i)\}_{i=1}^m$, then the weight list ρ can be formulated as :

$$\rho \leftarrow 1 - \{\frac{d_i}{\sum_{i=1}^m d_i}\}. \quad (5)$$

Regularization with Global Model

Early work (Zhang et al. 2019; Cui et al. 2021) claims that there exists a trade-off between accuracy and robustness, standard adversarial training can hurt accuracy. To achieve a lower boundary error A_{bdy}, we take advantage of logits from the global model f^{glo}, which is trained after aggregation. Particularly, in federated learning, the model owns the information obtained from the averaged parameters on distributed clients.

Let f^{loc} denote the adversarially trained model at each local client, f^{glo} has the most desirable classifier boundary for natural data. Then we can modify the local objective mentioned in Equation 3 as below:

$$\min \underbrace{\ell_{ce}(\rho \cdot f^{loc}(x^{adv}), y)}_{\text{for robustness}} + \beta \cdot \underbrace{\ell_{kl}(f^{loc}(x^{adv}), f^{glo}(x))}_{\text{for accuracy regularization}}. \quad (6)$$

Where ℓ_{ce} denotes the cross-entropy loss to improve the robustness, and ℓ_{kl} is the KL divergence loss to constrain the logits of global model and local model. Here, ℓ_{kl} appears as an additional regularization term, which is designed to reduce the boundary error $A_{bdy} = A_{cln} - A_{rob}$. Additionally, ρ is the weight calculated by Equation 5, β is the parameter to be tuned.

To show the difference between our DBFAT and existing defense methods, we list the loss functions of different adversarial training methods in Table 3.

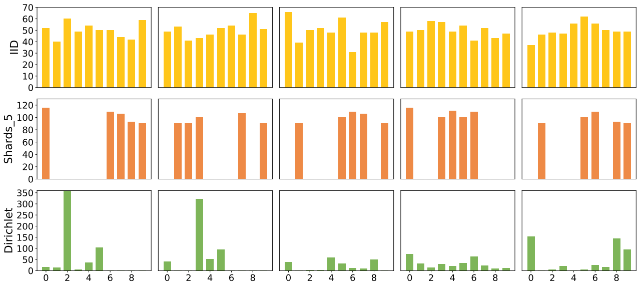

Figure 5: Visualizations of IID and non-IID distribution (Dirichlet sampled and Sharding) across 5 clients on CIFAR10 dataset. Shards_5 is a type of non-IID setting, in which each client has five categories of data (McMahan et al. 2017). From left to right: client ID number #1-5.

Experimental Results

Experimental Setup

Following the previous work of FL (McMahan et al. 2017), we distribute training data among 100 clients in both IID and non-IID fashion. For each communication round, we randomly select 10 clients to average the model parameters. All experiments are conducted with 8 Tesla V100 GPUs. More details can be referred to the supplemental material.

Datasets In this section, we show that DBFAT improves the robust generalization and meanwhile maintains a high accuracy with extensive experiments on benchmark CV datasets, including MNIST (Lecun et al. 1998), FashionMNIST (Xiao, Rasul, and Vollgraf 2017) (FMNIST), CIFAR10 (Krizhevsky and Hinton 2009), CIFAR100 (Krizhevsky and Hinton 2009), Tiny-ImageNet (Le and Yang 2015), and ImageNet-12 (Deng et al. 2009). The ImageNet-12 is generated via (Li et al. 2021c), which consists of 12 classes. We resize the original image with size 224*224*3 to 64*64*3 for fast training.

Data partitioning In the federated learning setup, we evaluate all algorithms on two types of non-IID data partitioning: **Dirichlet sampled data** and **Sharding**. For Dirichlet sampled data, each local client is allocated with a proportion of the samples of each label according to Dirichlet distribution (Li et al. 2020). Specifically, we follow the setting in (Yurochkin et al. 2019), for each label c, we sample $p_c \sim \mathrm{Dir}_J(0.5)$ and allocate $p_{c,j}$ proportion of the whole dataset of label c to client j. In this setting, some clients may entirely have no examples of a subset of classes. For Sharding (McMahan et al. 2017), each client owns data samples of a fixed number of labels. Let K be the number of total clients, and q is the number of labels we assign to each client. We divide the dataset by label into $K * q$ shards, and the amount of samples in each shard is $\frac{n}{K \cdot q}$. We denote this distribution as shards_q, where q controls the level of difficulty. If q is set to a smaller value, then the partition is more unbalanced. An example of these partitioning strategies is shown in Fig. 5, in which we visualize IID and non-IID distribution (Dirichlet sampled with $p_c \sim \mathrm{Dir}_J(0.5)$ and Sharding with shards_5) on five randomly selected clients.

Table 4: Accuracy and adversarial robustness on MNIST, FMNIST and CIFAR10 under both IID and non-IID distribution. An empirical study of FedAvg combined with several defense methods, more detailed comparisons are reported in the supplementary (Section B). Our method significantly outperforms other baselines.

Type		IID						Non-IID					
Dataset	Method	Clean	FGSM	MIM	PGD-20	CW	AA	Clean	FGSM	MIM	PGD-20	CW	AA
MNIST	Plain	99.01	28.35	8.65	5.29	3.84	3.02	98.45	11.78	14.06	8.44	9.51	7.45
	PGD_AT	98.52	76.01	60.18	54.50	55.23	50.43	97.82	67.58	52.89	48.03	47.43	43.75
	ALP	98.46	57.37	55.61	48.74	51.17	44.25	97.92	46.49	51.01	46.41	46.24	41.95
	TRADES	97.89	76.79	63.29	58.25	57.24	53.72	92.03	48.45	51.56	47.21	45.81	42.36
	AVMixup	98.63	61.41	53.34	42.33	46.95	37.78	97.47	56.50	51.86	46.28	44.46	41.84
	Ours	**98.86**	**78.06**	**70.97**	**68.39**	**63.09**	**59.39**	**97.95**	**68.54**	**54.18**	**50.33**	**49.12**	**44.32**
FMNIST	Plain	88.50	17.89	3.55	2.57	0.40	0.17	84.60	17.86	3.25	2.93	3.05	-1.40
	PGD_AT	76.05	68.53	65.24	65.40	64.26	60.89	72.93	60.11	54.42	54.33	52.19	49.88
	ALP	75.99	67.31	63.66	63.79	61.55	59.19	75.34	57.67	53.37	55.11	51.12	51.04
	TRADES	78.13	59.33	52.65	52.78	51.44	48.78	74.93	56.53	44.01	44.01	31.80	39.61
	AVMixup	79.34	61.22	54.93	54.67	49.48	50.07	72.06	56.26	49.21	49.72	47.99	45.15
	Ours	**81.49**	**69.23**	**66.22**	**66.24**	**65.71**	**61.49**	**76.19**	**63.11**	**56.45**	**58.31**	**56.96**	**53.91**
CIFAR10	Plain	78.80	6.87	1.15	1.06	1.30	1.23	61.10	7.58	2.94	2.67	2.87	1.28
	PGD_AT	58.75	30.62	27.23	26.11	28.47	22.09	15.27	13.27	13.00	13.00	12.99	8.63
	ALP	63.23	29.42	26.75	28.49	28.13	23.97	32.91	21.41	20.26	20.19	17.74	15.83
	TRADES	68.58	31.53	25.92	25.49	23.07	20.89	46.30	24.81	22.20	22.05	19.59	17.85
	AVMixup	70.28	29.51	26.22	26.34	24.07	22.25	48.23	25.29	21.42	24.25	20.25	19.43
	Ours	**72.21**	**31.47**	**28.57**	**29.03**	**29.31**	**24.25**	**52.24**	**27.03**	**24.12**	**27.02**	**22.13**	**21.20**

MNIST and FMNIST setup We use a simple CNN with two convolutional layers, followed by two fully connected layers. Following the setting used in (Goodfellow, Shlens, and Szegedy 2014), for MNIST, we set perturbation bound $\epsilon = 0.3$, and step size $\alpha = 0.01$, and apply adversarial attacks for 20 iterations. For FMNIST, we set perturbation bound $\epsilon = 32/255$, and step size $\alpha = 0.031$, we adversarially train the network for 10 steps and apply adversarial attacks for 20 iterations. Due to the simplicity of MNIST and FMNIST, we mainly use non-IID data (Sharding), which is hard to train.

CIFAR10, CIFAR100, Tiny-ImageNet and ImageNet-12 setup We apply a larger CNN architecture, and follow the setting used in (Madry et al. 2017), i.e., we set the perturbation bound $\epsilon = 0.031$, step size $\alpha = 0.007$. To evaluate the robustness, we conduct extensive experiments with various data partitioning.

Baselines For attack methods, we perform five popular attacks including FGSM (Kurakin, Goodfellow, and Bengio 2017), MIM (Dong et al. 2018), PGD (Madry et al. 2017), CW (Carlini and Wagner 2017) and AA (Croce and Hein 2020). We further use Square (Andriushchenko et al. 2020) for black-box attack. To investigate the effectiveness of existing FL algorithms, we implement FedAvg(McMahan et al. 2017), FedProx(Li et al. 2018), FedNova(Wang et al. 2020) and Scaffold(Karimireddy et al. 2020). To defend against adversarial attacks, we implement four most prevailing methods including PGD_AT(Madry et al. 2017), TRADES (Zhang et al. 2019), ALP (Kannan, Kurakin, and Goodfellow 2018), MMA (Ding et al. 2020) and AVMixup (Lee, Lee, and Yoon 2020). We compare the performance of our DBFAT with various kinds of defense methods combined with FL methods.

Convergence For Local Training

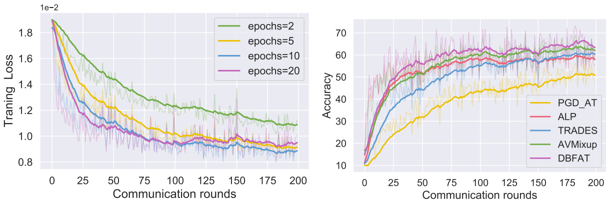

Figure 6: **Left:** Convergence rate for different local epochs. **Right:** Training curves of FedAvg combined with different AT methods.

To show the convergence rate of DBFAT, we use the Dirichlet sampled CIFAR10 dataset, where each client owns 500 samples from 5 classes. Fig. 6 (left sub-figure) shows the impact of local epoch E during adversarial training. Indeed, for a very small epoch (e.g., $E = 2$), it has an extremely slow convergence rate, which may incur more communications. Besides, a large epoch (e.g., $E = 20$) also leads to a slow convergence, as model may overfit to the local data. Considering both the communication cost and convergence issues, we set $E = 5$ in our experiments, which can maintain a proper communication efficiency and fast convergence.

Effectiveness of Our Method

We verify the effectiveness of our method compared with several adversarial training techniques on Dirichlet sampled CIFAR10. Evaluation of model robustness is averaged under four attacks using the the same setting for a fair comparison and all defense methods are combined with FedAvg.

To show the differences between DBFAT and above mentioned defense methods, we report the training curves on

Table 5: Accuracy and adversarial robustness on CIFAR100, Tiny-ImageNet, and ImageNet-12.

Dataset	CIFAR100				Tiny-ImageNet				ImageNet-12			
Method	Clean	PGD-20	AA	Square	Clean	PGD-20	AA	Square	Clean	PGD-20	AA	Square
PGD_AT	39.32	16.07	14.36	23.44	26.33	12.26	10.26	13.54	37.42	22.61	18.30	25.57
ALP	41.12	18.46	14.78	24.54	32.78	14.62	12.19	16.48	54.96	24.78	19.57	27.73
TRADES	43.39	20.05	16.85	26.43	37.81	15.49	13.26	19.38	58.82	25.49	21.81	28.96
AVMixup	46.64	23.56	19.46	29.16	36.19	15.28	13.18	19.25	59.63	25.81	21.92	29.28
Ours	**48.31**	**24.47**	**22.46**	**31.57**	**38.24**	**16.17**	**13.96**	**20.26**	**61.38**	**26.47**	**22.08**	**30.91**

Table 6: Ablation Study by cutting off different modules.

Dataset	CIFAR10		FMNIST	
Methods	A_{cln}	Avg A_{rob}	A_{cln}	Avg A_{rob}
Ours	**52.16**	**27.80**	**75.89**	**59.63**
Ours (w/o re-weighting)	48.44	25.89	72.35	56.34
Ours (w/o regularization)	51.04	26.84	73.96	58.23

Table 7: Effect of hyper-parameter β. "Avg A_{rob}" refers to the average robustness under four attacks.

Dataset	MNIST		FMNIST	
β	A_{cln}	Avg A_{rob}	A_{cln}	Avg A_{rob}
4	98.30	26.64	81.73	37.36
2	98.14	34.24	**75.59**	**47.83**
1.5	**98.46**	**53.22**	74.93	44.08
1	97.32	47.35	65.43	42.33
0.5	96.57	44.09	61.02	45.28

non-IID CIFAR10 dataset in the right sub-figure of Fig. 6. Fig. 6 confirms that our DBFAT achieves the highest clean accuracy. We speculate that this benefit is due to the regularization term and re-weighting strategy introduced in Equation 6. It is worth mentioning that in the training curves, the model trained with PGD_AT performs very poorly. It indicates that standard AT may not be a suitable choice for adversarial robustness in FL, as it only uses cross-entropy loss with adversarial examples, but ignores the negative impact on clean accuracy. We further report the results on various datasets under both IID and non-IID settings in Table 4, which indicates that DBFAT significantly outperforms other methods in terms of both accuracy and robustness.

Performance on large datasets In Table 5, we show the accuracy and robustness of each method on large datasets (e.g., CIFAR100, Tiny-ImageNet, and ImageNet-12). All results are tested under PGD-20 attack (Madry et al. 2017), AutoAttack (Croce and Hein 2020), and Square attack (Andriushchenko et al. 2020) in non-IID settings. From the results reported in Table 5, we can find that our method still outperforms other baselines in terms of both clean accuracy and robustness. Note that our method can achieve the highest accuracy and robustness of 61.38% and 22.08% under AutoAttack, respectively. It thus proves that our method can also be used to improve the accuracy and robustness of the model on large datasets. We think that the higher clean accuracy is a result of the regularization term introduced in Equation 6, while maintaining a high robustness.

Ablation Study

Cutting off different modules As part of our ablation study, we first investigate the contributions of different modules introduced in DBFAT. As shown in Table 6, turning off both the re-weighting strategy and regularization term will lead to poor performance, which demonstrates the importance of both modules. Moreover, cut-offing the re-weighting strategy can lead to a more severe degradation. We conjecture this is a reasonable phenomenon. As mentioned in Fig. 1, non-IID data can cause a serious accuracy reduction. Our re-weighting strategy can alleviate the bias by taking the limited data on each client into account.

Effects of Regularization The regularization parameter β is an important hyperparameter in our proposed method. We show how the regularization parameter affects the performance of our robust classifiers by numerical experiments on two datasets, MNIST and FMNIST. In Equation 6, β controls the accuracy obtained from the global model, which contains information from distributed clients. Since directly training on adversarial examples could hurt the clean accuracy, here we explore the effects of β on both accuracy and robustness. As shown in Table 7, we report the clean accuracy and robustness by varying the value of β. We empirically choose the best β for different datasets. For example, for MNIST, $\beta = 1.5$ can achieve better accuracy and robustness. For FMNIST, we let $\beta = 2$ for a proper trade-off in accuracy and robustness.

Conclusion

In this paper, we investigate an interesting yet not well explored problem in FL: the robustness against adversarial attacks. We first find that directly adopting adversarial training in federated learning can hurt accuracy significantly especially in non-IID setting. We then propose a novel and effective adversarial training method called DBFAT, which is based on the decision boundary of federated learning, and utilizes local re-weighting and global regularization to improve both accuracy and robustness of FL systems. Comprehensive experiments on various datasets and detailed comparisons with the state-of-the-art adversarial training methods demonstrate that our proposed DBFAT consistently outperforms other baselines under both IID and non-IID settings. This work would potentially benefit researchers who are interested in adversarial robustness of FL.

References

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square Attack: a query-efficient black-box adversarial attack via random search. arXiv:1912.00049.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Chen, C.; Kailkhura, B.; Goldhahn, R.; and Zhou, Y. 2021a. Certifiably-Robust Federated Adversarial Learning via Randomized Smoothing. arXiv:2103.16031.

Chen, C.; Liu, Y.; Ma, X.; and Lyu, L. 2022a. CalFAT: Calibrated Federated Adversarial Training with Label Skewness. In *Advances in Neural Information Processing Systems*.

Chen, C.; Zhang, J.; and Lyu, L. 2022. Gear: a margin-based federated adversarial training approach. In *International Workshop on Trustable, Verifiable, and Auditable Federated Learning in Conjunction with AAAI*, volume 2022.

Chen, Z.; Li, B.; Wu, S.; Xu, J.; Ding, S.; and Zhang, W. 2022b. Shape Matters: Deformable Patch Attack. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022*. Springer.

Chen, Z.; Li, B.; Xu, J.; Wu, S.; Ding, S.; and Zhang, W. 2022c. Towards Practical Certifiable Patch Defense With Vision Transformer. In *Proceedings of the IEEE/CVF Conference on CVPR*, 15148–15158.

Chen, Z.; Zhu, M.; Yang, C.; and Yuan, Y. 2021b. Personalized Retrogress-Resilient Framework for Real-World Medical Federated Learning. In de Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; and Essert, C., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part III*, volume 12903 of *Lecture Notes in Computer Science*, 347–356. Springer.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. arXiv:2003.01690.

Cui, J.; Liu, S.; Wang, L.; and Jia, J. 2021. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15721–15730.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2020. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. arXiv:1812.02637.

Dong, J.; Cong, Y.; Sun, G.; Fang, Z.; and Ding, Z. 2021. Where and How to Transfer: Knowledge Aggregation-Induced Transferability Perception for Unsupervised Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(1): 1–17.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2018. Knowledge Distillation with Adversarial Samples Supporting Decision Boundary. arXiv:1805.05532.

Hong, J.; Wang, H.; Wang, Z.; and Zhou, J. 2021. Federated Robustness Propagation: Sharing Adversarial Robustness in Federated Learning. *arXiv preprint arXiv:2106.10196*.

Huang, R.; Cui, C.; Chen, F.; Ren, Y.; Liu, J.; Zhao, Z.; Huai, B.; and Wang, Z. 2022a. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2525–2535.

Huang, R.; Lam, M. W.; Wang, J.; Su, D.; Yu, D.; Ren, Y.; and Zhao, Z. 2022b. FastDiff: A Fast Conditional Diffusion Model for High-Quality Speech Synthesis. *arXiv preprint arXiv:2204.09934*.

Kairouz, P. 2021. Advances and Open Problems in Federated Learning. arXiv:1912.04977.

Kannan, H.; Kurakin, A.; and Goodfellow, I. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017. Adversarial examples in the physical world. arXiv:1607.02533.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, S.; Lee, H.; and Yoon, S. 2020. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 272–281.

Li, B.; Sun, Z.; and Guo, Y. 2019. SuperVAE: Superpixelwise Variational Autoencoder for Salient Object Detection. In *The Thirty-Third AAAI Conference*.

Li, B.; Sun, Z.; Tang, L.; Sun, Y.; and Shi, J. 2019. Detecting Robust Co-Saliency with Recurrent Co-Attention Neural Network. In Kraus, S., ed., *IJCAI*.

Li, B.; Xu, J.; Wu, S.; Ding, S.; Li, J.; and Huang, F. 2021a. Detecting Adversarial Patch Attacks through Global-local Consistency. In *ADVM '21: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 35–41. ACM.

Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2021b. Federated Learning on Non-IID Data Silos: An Experimental Study. *arXiv preprint arXiv:2102.02079*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020. On the Convergence of FedAvg on Non-IID Data. arXiv:1907.02189.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021c. Anti-backdoor learning: Training clean models on poisoned data. *NeurIPS*, 34.

Liang, F.; Pan, W.; and Ming, Z. 2021. FedRec++: Lossless Federated Recommendation with Explicit Feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 4224–4231. AAAI Press.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P. 2021a. FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 1013–1023. Computer Vision Foundation / IEEE.

Liu, S.; Xu, S.; Yu, W.; Fu, Z.; Zhang, Y.; and Marian, A. 2021b. FedCT: Federated Collaborative Transfer for Recommendation. In Diaz, F.; Shah, C.; Suel, T.; Castells, P.; Jones, R.; and Sakai, T., eds., *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 716–725. ACM.

Lyu, L.; Yu, H.; Ma, X.; Chen, C.; Sun, L.; Zhao, J.; Yang, Q.; and Philip, S. Y. 2022. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

Shah, D.; Dube, P.; Chakraborty, S.; and Verma, A. 2021. Adversarial training in communication constrained federated learning. *arXiv preprint arXiv:2103.01319*.

Wang, C.; Deng, J.; Meng, X.; Wang, Y.; Li, J.; Miao, F.; Rajasekaran, S.; and Ding, C. 2021. A Secure and Efficient Federated Learning Framework for NLP. In *EMNLP 2021*, 7676–7682. Association for Computational Linguistics.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747.

Yurochkin, M.; Agarwal, M.; Ghosh, S.; Greenewald, K.; Hoang, T. N.; and Khazaeni, Y. 2019. Bayesian Nonparametric Federated Learning of Neural Networks. arXiv:1905.12022.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 7472–7482. PMLR.

Zhang, J.; Chen, C.; Li, B.; Lyu, L.; Wu, S.; Ding, S.; Shen, C.; and Wu, C. 2022a. DENSE: Data-Free One-Shot Federated Learning. In *Advances in NeurIPS*.

Zhang, J.; Li, B.; Xu, J.; Wu, S.; Ding, S.; Zhang, L.; and Wu, C. 2022b. Towards Efficient Data Free Black-Box Adversarial Attack. In *Proceedings of the IEEE/CVF Conference on CVPR*, 15115–15125.

Zhang, J.; Li, Z.; Li, B.; Xu, J.; Wu, S.; Ding, S.; and Wu, C. 2022c. Federated Learning with Label Distribution Skew via Logits Calibration. In *Proceedings of the ICML*. PMLR.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2021. Geometry-aware Instance-reweighted Adversarial Training. In *International Conference on Learning Representations*.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated Learning with Non-IID Data. arXiv:1806.00582.

Zhou, Y.; Wu, J.; Wang, H.; and He, J. 2022. Adversarial robustness through bias variance decomposition: A new perspective for federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2753–2762.

Zhu, X.; Wang, J.; Hong, Z.; and Xiao, J. 2020. Empirical Studies of Institutional Federated Learning For Natural Language Processing. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, 625–634. Association for Computational Linguistics.

Zizzo, G.; Rawat, A.; Sinn, M.; and Buesser, B. 2020. FAT: Federated Adversarial Training. arXiv:2012.01791.