FTL-IJCAI'21

Ensemble Federated Adversarial Training with Non-IID data

Shuang Luo, Didi Zhu, Zexi Li, Chao Wu

Zhejiang University, Hangzhou 310007, China

Abstract

Despite federated learning endows distributed clients with a cooperative training mode under the premise of protecting data privacy and security, the clients are still vulnerable when encountering adversarial samples due to the lack of robustness. The adversarial samples can confuse and cheat the client models to achieve malicious purposes via injecting elaborate noise into normal input. In this paper, we introduce a novel Ensemble Federated Adversarial Training Method, termed as EFAT, that enables an efficacious and robust coupled training mechanism. Our core idea is to enhance the diversity of adversarial examples through expanding training data with different disturbances generated from other participated clients, which helps adversarial training perform well in Non-IID settings. Experimental results on different Non-IID situations, including feature distribution skew and label distribution skew, show that our proposed method achieves promising results compared with solely combining federated learning with adversarial approaches.

Contributions

We propose the Ensemble Federated Adversarial Training (EFAT) method to improve the robustness of models against black-box attacks with non-IID training data by taking advantage of improving adversarial data diversity between models from distributed clients.

- We explore the impact of adversarial training on the federated training paradigm and find it plays an important role. To this end, we develop a novel ensemble federated adversarial training (EFAT) methodology by incorporating adversarial examples generated by other clients' models to improve each client's robustness.
- Building on the above insight, we demonstrate our methodology's effectiveness and robustness against black-box attacks during inference-time on two kinds of Non-IID settings, including feature distribution skew and label distribution skew. The evaluation result shows that EFAT reaches higher adversarial accuracy on both Digit-Five and CIFAR10 than baseline.

Ensemble Federated Adversarial Training Methods

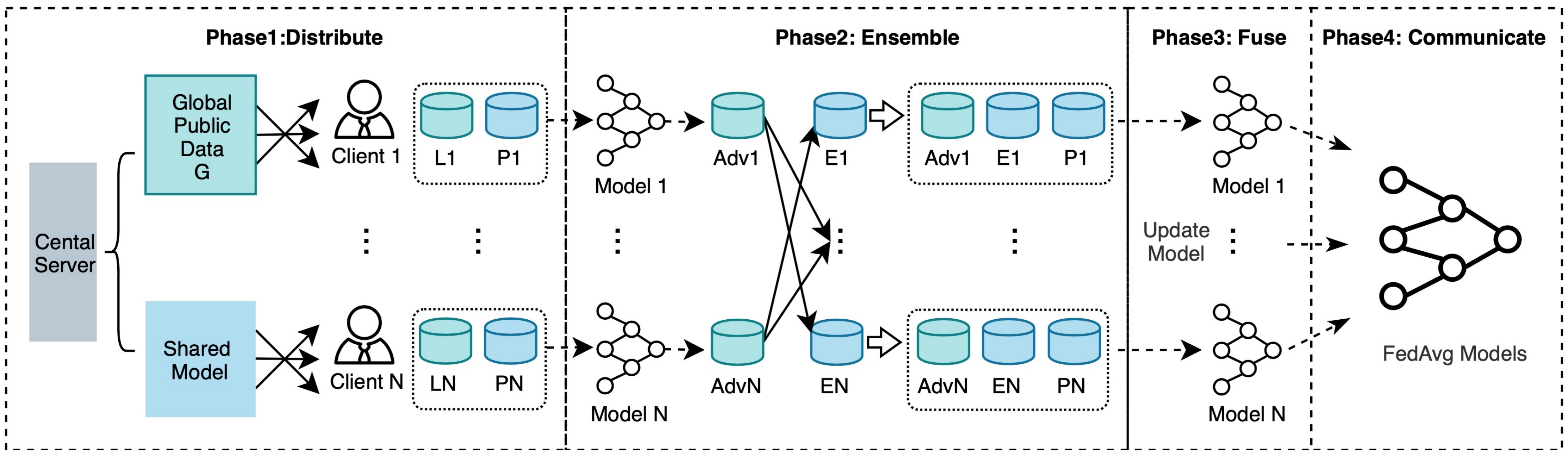

Fig 1. Illustration of the proposed EFAT framework.

The EFAT method involves 4 phases:

(1) Distribute: Distributing the shared model and parts of global public dataset L to all the clients. P is the clients' private data.

(2) Ensemble: Integrating adversarial samples Adv generated from the local public dataset L of other clients to form ensemble training data E.

(3) Fuse: Fusing the various data distribution including the potential knowledge of other clients by adversarial training.

(4) Communicate: Client model updates are aggregated on the central server using the FedAvg algorithm.

Experiments

non-i.i.d.-ness	IID			Non-IID					
	γ=100			γ=1			γ=0.01		
Method	clean	PGD-10	PGD-20	clean	PGD-10	PGD-20	clean	PGD-10	PGD-20
Baseline	72.21%	62.34%	64.17%	72.21%	63.02%	63.62%	72.46%	65.05%	64.62%
EFNT	**80.45%**	43.91%	41.02%	**79.03%**	43.23%	43.99%	**81.19%**	47.82%	42.79%
EFNT+AT	78.42%	57.45%	48.35%	78.81%	58.36%	46.41%	81.15%	56.24%	49.41%
EFAT	72.02%	**70.83%**	**67.25%**	72.64%	**70.28%**	**68.64%**	74.66%	**71.48%**	**67.46%**

Table 2. Accuracies of CIFAR10 under black-box attacks in IID and non-IID settings w.r.t. α = 5%.

non-i.i.d.-ness	IID			Non-IID					
	γ=100			γ=1			γ=0.01		
Method	clean	PGD-10	PGD-20	clean	PGD-10	PGD-20	clean	PGD-10	PGD-20
Baseline	72.91%	62.46%	61.79%	72.95%	61.83%	60.41%	72.93%	61.29%	59.80%
EFNT	**82.32%**	43.51%	43.49%	**80.23%**	50.17%	44.84%	**82.89%**	45.36%	42.20%
EFNT+AT	81.45%	65.26%	62.45%	79.54%	64.65%	62.12%	81.57%	65.04%	62.12%
EFAT	73.39%	**70.47%**	**68.35%**	73.45%	**70.98%**	**68.25%**	75.57%	**71.66%**	**68.43%**

Table 2. Accuracies of CIFAR10 under black-box attacks in IID and non-IID settings w.r.t. α = 10%.

We use Digit-Five datasets as feature distribution skew datasets, which is a collection of five benchmarks for digit recognition and construct a label distributed skew dataset based on CIFAR10 by using the Dirichlet distribution. The value of γ controls the degree of non-i.i.d.-ness.

The results of experiments show that compared with the models only trained locally with their own adversarial examples(baseline), EFAT, EFNT, and EFNT+AT trained with exchange public data reach higher accuracy against attacks. It is because in the setting of feature distribution skew, expanding training data from different clients' models increases the diversity of training data distribution, which helps improve the robustness of models.

Method	MNIST,SVHN, MNISTM,USPS →SYN	MNIST,SVHN, MNISTM,SYN →USPS	MNIST,SVHN, USPS,SYN →MNISTM	MNIST,USPS, MNISTM,SYN →SVHN	MNIST,SVHN, MNISTM,SYN →MNIST
Baseline	61.78%	78.06%	58.61%	27.15%	90.05%
EFNT	78.06%	82.50%	**73.60%**	33.75%	98.26%
EFNT+AT	82.30%	82.35%	**73.60%**	**48.80%**	98.55%
EFAT	**85.84%**	**83.45%**	71.65%	45.50%	**98.65%**

Table 1. Performance of clients trained with Digit-Five against black-box PGD adversaries.

Conclusion

In this paper we present a novel ensemble federated adversarial training method, termed as EFAT, to improve the robust- ness of models against black-box attacks in federated learning. The proposed method enhances the diversity of adversarial examples through expanding training data with perturbations generated from other participating clients. Experiment results on both Digit-Five and CIFAR10 in IID and Non-IID settings show that our method significantly im- proves the robustness and accuracy contrasted with the intuitive federated adversarial training method and the other two variants of EFAT.

Please contact us if you are interested in our work!
Email: luoshuang@zju.edu.cn, chao.wu@zju.edu.cn