Multi-metrics adaptively identifies backdoors in Federated learning

Siquan Huang[1] Yijiang Li[2] Chong Chen[1] Leyu Shi[1] Ying Gao[1]
[1]South China University of Technology [2]Johns Hopkins University

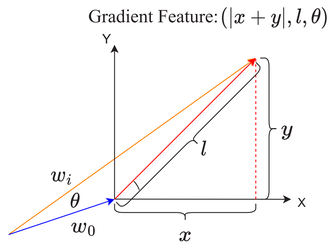
ICCV23 PARIS

Code
https://github.com/siquanhuang/Multi-metrics_against_backdoors_in_FL

Email
cssiquanhuang@mail.scut.edu.cn

Problem and Contribution

Problem: 1. Euclidean distance (ie, L_2 distance) suffers from the curse of dimensionality. 2. Single metric takes effect only against particular attacks with detailed assumptions regarding the malicious gradients.

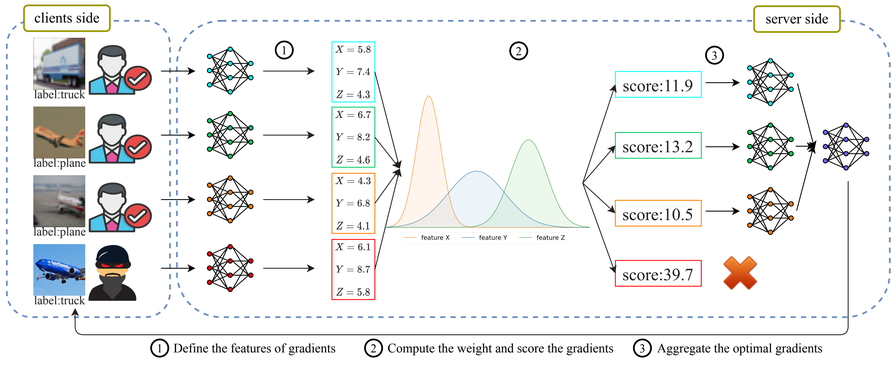


Contributions:

- We present a novel defense with multi-metrics to adaptively identify backdoors, which is applicable in a generic adversary model without predefined assumptions over the attack strategy or data distribution.
- We show that by introducing the Manhattan distance, our defense alleviates the "meaningfulness" problem of Euclidean distance in high dimensions.
- By utilizing multiple metrics with dynamic weighting, our defense can resist backdoor attacks under various attack settings and data distributions.

The Proposed Method

Method Framework:



The Superiority of Manhattan Distance:

- Manhattan can discriminate more than the Euclidean distance in high-dimension space.

Theorem:

$$\text{If } \lim_{d\to\infty} var\left(\frac{\|X_d\|_k}{E\left[\|X_d\|_k\right]}\right) = 0, \text{then}$$

$$\frac{Dmax_d^k - Dmin_d^k}{D\min_d^k} \to_p 0,$$

Lemma:

$$\lim_{d\to\infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{(1/k)-(1/2)}}\right] = C_k,$$

Proposition:

$$\lim_{d\to\infty} E\left[\frac{M_d}{U_d \cdot d^{\frac{1}{2}}}\right] = C'.$$

Dynamic Weighting:

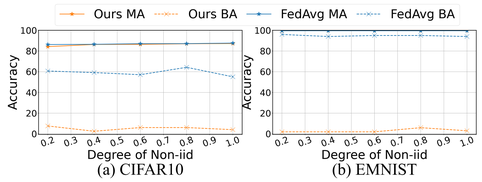
$$\delta^{(i)} = \sqrt{\boldsymbol{x'}^{(i)\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{x'}^{(i)}}.$$

Experiments & Results

Robustness against Different Attacks and Comparison with SOTA

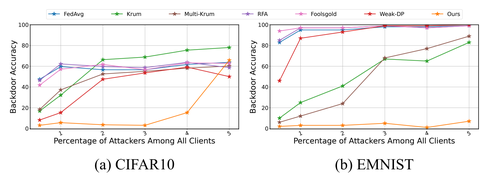
Dataset	Defense	Model Replacement [4]		DBA [48]		PGD [45]		Edge-case PGD [45]		Ranking Score ↑
		MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓	
CIFAR10	FedAvg [32]	**86.95**	64.80	79.23	90.44	87.04	14.44	87.14	55.10	0
	RFA [38]	86.69(+0.00)	25.56(-0.61)	79.6(+0.00)	57.69(-0.36)	**87.1**(+0.00)	52.56(+2.64)	86.47(-0.01)	65.31(+0.19)	-1.86
	Foolsgold [16]	85.71(-0.01)	6.67(-0.90)	77.56(-0.02)	3.43(-0.96)	84.92(-0.02)	14.44(+0.00)	85.72(-0.02)	45.41(-0.18)	+1.96
	Krum [8]	82.17(-0.05)	6.11(-0.91)	78.18(-0.01)	6.01(-0.93)	82.32(-0.05)	66.67(+3.62)	81.23(-0.07)	59.18(+0.07)	-2.04
	Multi-Krum [8]	86.55(+0.00)	1.67(-0.97)	79.33(0.00)	91.39(+0.01)	86.52(-0.01)	17.78(+0.23)	**87.4**(+0.00)	60.2(+0.09)	+0.63
	Weak-DP [44]	74.41(-0.14)	46.11(-0.29)	10.00(-0.87)	**0.00**(-1.00)	73.61(-0.15)	12.78(-0.11)	73.84(-0.15)	53.06(-0.04)	+0.12
	Flame [35]	80.58(-0.07)	**0.56**(-0.99)	76.78(-0.03)	37.24(-0.59)	81.24(-0.07)	**0.56**(-0.96)	81.41(-0.07)	5.12(-0.91)	+3.21
	Ours	86.34(-0.01)	**0.56**(-0.99)	**79.61**(+0.00)	9.98(-0.89)	86.44(-0.01)	0.56(-0.96)	86.86(+0.00)	**3.06**(-0.94)	**+3.77**
EMNIST	FedAvg [32]	99.54	96.00	97.68	94.13	**99.55**	10.00	99.37	96.00	0
	RFA [38]	99.57(+0.00)	6.00(-0.94)	**97.87**(+0.00)	1.39(-0.99)	99.32(+0.00)	4.00(-0.60)	99.29(+0.00)	97.00(+0.01)	+2.51
	Foolsgold [16]	96.42(-0.03)	98.00(+0.02)	97.24(+0.00)	0.64(-0.99)	99.07(+0.00)	94.00(+8.40)	99.13(+0.00)	98.00(+0.02)	-7.49
	Krum [8]	99.22(+0.00)	**0.00**(-1.00)	97.7(+0.00)	0.56(-0.99)	99.12(+0.00)	1.00(-0.90)	99.14(+0.00)	12.00(-0.88)	+3.76
	Multi-Krum [8]	**99.58**(+0.00)	**0.00**(-1.00)	97.85(+0.00)	47.43(-0.50)	99.54(+0.00)	0.00(-1.00)	99.57(+0.00)	84.00(-0.13)	+2.63
	Weak-DP [44]	99.37(+0.00)	86.00(-0.10)	10.00(-0.90)	**0.00**(-1.00)	99.41(+0.00)	14.00(+0.40)	99.39(+0.00)	89.00(-0.07)	-0.12
	Flame [35]	99.39(+0.00)	**0.00**(-1.00)	97.12(-0.01)	17.38(-0.82)	99.39(+0.00)	**0.00**(-1.00)	99.44(+0.00)	13.00(-0.86)	+3.67
	Ours	99.53(+0.00)	**0.00**(-1.00)	97.39(+0.00)	4.23(-0.96)	99.54(+0.00)	**0.00**(-1.00)	**99.58**(+0.00)	**0.00**(-1.00)	**+3.95**

Impact of Different Degrees of Non-IID



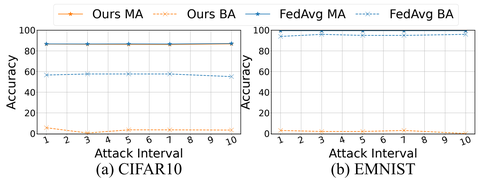
(a) CIFAR10 (b) EMNIST

Impact of Attacker Percentage



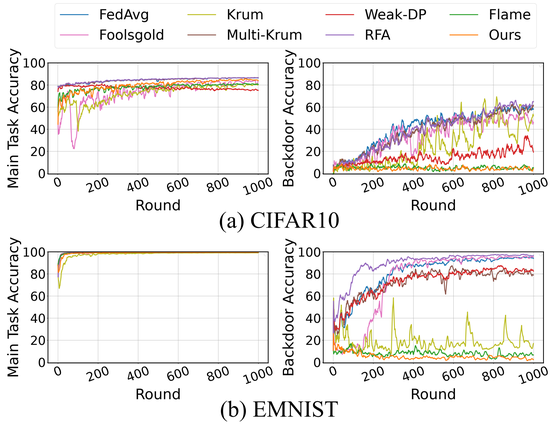
(a) CIFAR10 (b) EMNIST

Impact of Attack Frequency



(a) CIFAR10 (b) EMNIST

Training process under Edge-case PGD attack

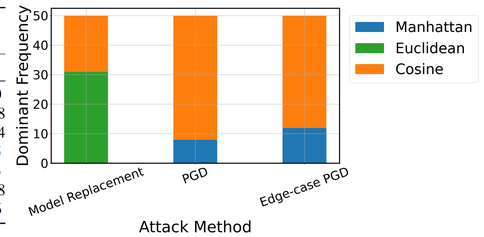


(a) CIFAR10

(b) EMNIST

Ablation Study on Metrics

Defenses	Model Replacement	PGD	Edge-case PGD
	MA/BA	MA/BA	MA/BA
Man	83.86/**0.56**	83.74/25.56	85.3/64.80
Eul	86.24/**0.56**	85.52/17.78	**87.12**/54.08
Cos	84.22/2.22	83.84/30.00	85.38/66.84
Man+Eul	84.09/1.11	84.17/28.63	84.3/67.35
Man+Cosine	85.74/1.67	85.16/23.68	85.86/6.63
Cosine+Eul	86.31/**0.56**	85.44/16.11	85.14/63.78
Man+Cosine+Eul	**86.34/0.56**	**86.44/0.56**	86.86/**3.06**



Ablation Study on Dynamic Weighting

Weighting	Model Replacement		PGD		Edge-case PGD	
	MA ↑	BA ↓	MA↑	BA ↓	MA ↑	BA ↓
Max Norm	83.86	**0.56**	83.74	25.56	84.08	62.24
Whitening	**86.34**	**0.56**	**86.44**	**0.56**	**86.86**	**3.06**

Generalization to Different Datasets

Defense	CINIC10		LOAN		Sentiment140	
	MA ↑	BA ↓	MA ↑	BA ↓	MA ↑	BA ↓
FedAvg	**80.02**	36.22	**89.05**	61.36	**82.59**	89.17
Ours	76.24	**4.59**	88.52	**0**	81.67	**5.83**

Conclusion

- By leveraging multiple metrics with dynamic weighting, the proposed multi-metrics defense withstand a wide range of stealthy and elaborate attacks in FL;
- The proposed method achieves state-of-the-art performance, especially against the Edge-case PGD attack.