



南京航空航天大学

Nanjing University of Aeronautics and Astronautics

南京航空航天大学

Nanjing University of Aeronautics and Astronautics



Mitigating Backdoor Attacks in Federated Learning

Chen Wu

Computer Science and Engineering

Pennsylvania State University

cvw5218@psu.edu

Xian Yang

Computer Science

North Carolina State University

xyang45@ncsu.edu

Sencun Zhu

Computer Science and Engineering

Pennsylvania State University

sxz16@psu.edu

Prasenjit Mitra

Information Sciences and Technology

Pennsylvania State University

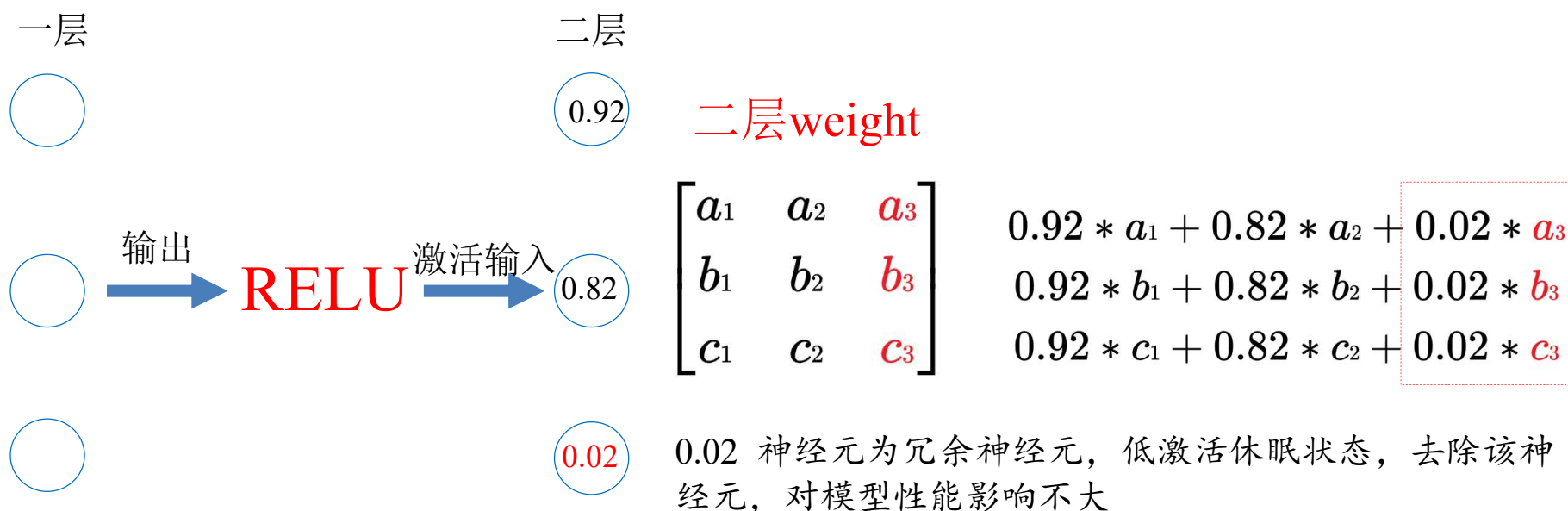
pum10@psu.edu

arxiv 2021

Backdoor Attack

模型剪枝(Model Pruning)

深度学习网络模型从卷积层到全连接层存在着大量冗余的参数，大量神经元激活值趋近于0，将这些神经元去除后可以表现出同样的模型表达能力，这种情况被称为过参数化，而对应的技术则被称为模型剪枝。

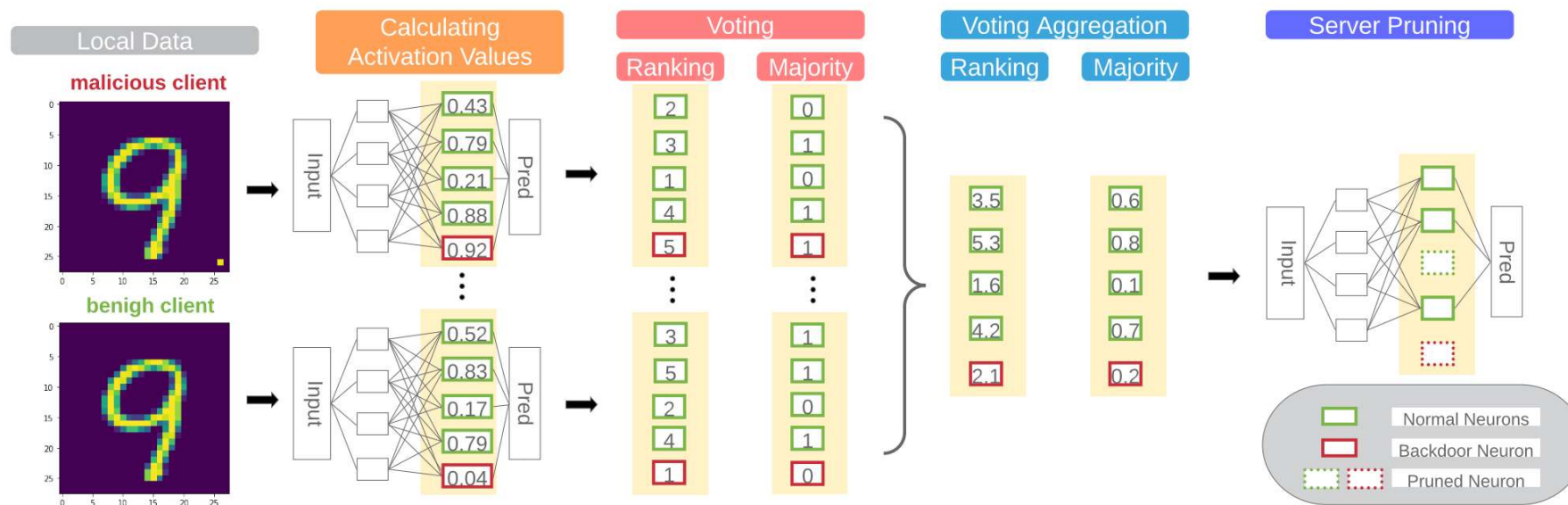


剪枝：将与该神经元关联的权重列设置为0 （列剪枝）

Backdoor Attack

BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

BadNets发现，带trigger(触发器)样本输入后门模型，会让模型产生一些区别于良性输入的差异。
输入触发器样本，后门模型会激活一些所谓“后门神经元”，而这些神经元在面对良性输入时，常处于休眠状态(低激活)。





Backdoor Attack

后门防御评估指标

1 .Attack Success Rate(ASR)

攻击成功率，输入触发器样本，模型识别为target class(攻击类别)成功率，ASR越大，模型后门性能越强

2. Main Task Accuracy(ACC)

主任务(良性任务)，在非触发样本上的性能，ACC 值越大，模型的预测性能越强。

攻击者目标：ASR和ACC同时最大。

防御者目标：ASR最小和ACC最大



Backdoor Attack

剪枝后门防御原理

去除休眠冗余神经元，防止在后门攻击时这些休眠神经元被激活。

剪枝标准

休眠神经元是人为认定，过度剪枝会影响模型性能。一般使用辅助验证集，边剪枝边评估当前模型在良性样本上的准确度。当模型在辅助数据集上的准确度下降到阈值停止剪枝。

Mitigating Backdoor Attacks in Federated Learning

假设1. 拥有辅助数据集

服务器使用数据集去寻找休眠神经元。(优点：自身评估结果，可信。缺点：拥有验证数据集)

假设2. 无辅助数据集

借助客户端本地验证集去评估，客户端上传评估结果。(优点：服务器无需拥有验证数据集。缺点：无法排除后门客户端告诉的虚假序列)

EXPERIMENTS ON FASHION-MNIST DATASET

Target		Training Phase		Pruning Neurons		Adjusting Extreme Weights		Defense with Fine-tuning	
vic	atk	test acc	atk acc	test acc	atk acc	test acc	atk acc	test acc	atk acc
9	0	88.8	99.8	83.2	2.8	82.9	2.1	86.3	9.0
9	1	88.7	99.4	82.6	6.5	82.1	0	87.0	0
9	2	87.8	99.8	82.2	2.7	82.1	3.1	85.6	12.6
9	3	87.5	99.6	84.4	0.3	84.4	0	86.2	0.4
9	4	87.8	99.7	81.0	2.9	80.6	0	85.8	2.3
9	5	88.6	99.7	81.2	11.9	80.6	3.6	86.1	10.4
9	6	86.3	99.8	82.8	93.6	82.0	0.2	86.0	3.1
9	7	88.5	99.7	82.9	4.3	83.1	4.6	87.0	17.1
9	8	88.8	99.9	84.7	87.7	85.0	3.2	87.2	2.9

原始类别 攻击类别 ACC ASR 防御前 只剪枝 剪枝+极端权值调整 剪枝+参数微调

区别于神经元剪枝将休眠权重设置为0，将休眠神经元关联的权重列，在全局中训练得到的增量进行方向翻转。

```
for index in  $L_0^\tau$  do
     $w_T[:, index] \leftarrow 0$ 
```

原有剪枝($a_3, b_3, c_3 = 0$)
 a_3
 b_3 为低激活神经元关联权重列
 c_3

```
for index in  $L_0^\tau$  do
     $w_T[:, index] \leftarrow w_0[:, index] - update^\tau[:, index]$ 
```

```
for index in  $L_1^\tau$  do
     $w_T[:, index] \leftarrow w_0[:, index] + update^\tau[:, index]$ 
```

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$$

高激活值 依赖 关联权重 传递到 输出层 形成攻击效果。而这些关联权重在全局获得的 Updates 起到关键作用，Flip Updates。

Experiments

IID

Dataset		MNIST		FMNIST		EMNIST		CIFAR-10 _(DBA)	
Base	Target	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)
0	5	0.949	0.992	0.991	0.901	1	0.854	0.974	0.837
		0	0.992	0	0.900	0	0.850	0.019	0.821
	6	0.965	0.991	0.992	0.905	0.999	0.851	0.960	0.836
		0	0.990	0.036	0.880	0	0.847	0.022	0.820
	7	1	0.989	0.989	0.904	0.999	0.856	0.972	0.820
		0	0.989	0	0.894	0	0.856	0.007	0.790
1	5	0.999	0.989	0.994	0.903	0.999	0.857	0.911	0.827
		0	0.989	0.003	0.902	0	0.854	0.021	0.805
	6	0.999	0.988	0.996	0.903	0.999	0.862	0.912	0.807
		0	0.986	0.041	0.901	0	0.854	0.027	0.783
	7	1	0.989	0.996	0.903	0.996	0.846	0.929	0.840
		0	0.989	0	0.886	0.001	0.846	0.048	0.798

Experiments

non-IID
Dirichlet $a = 0.5$

Dataset		MNIST		FMNIST		EMNIST		CIFAR-10 _(CBA)	
Base	Target	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)	ASR(%)	ACC(%)
0	5	1	0.988	0.989	0.897	0.999	0.835	0.969	0.762
		0	0.988	0	0.880	0	0.832	0.002	0.754
	6	1	0.989	0.993	0.892	0.999	0.848	0.981	0.781
		0	0.989	0.006	0.880	0	0.847	0.002	0.772
	7	1	0.989	0.983	0.883	0.997	0.831	0.961	0.772
		0	0.988	0.001	0.869	0	0.827	0.016	0.761
1	5	0.999	0.989	0.977	0.886	0.999	0.838	0.935	0.765
		0	0.989	0	0.862	0	0.837	0.008	0.730
	6	0.999	0.987	0.993	0.886	0.997	0.837	0.933	0.755
		0	0.987	0.002	0.877	0	0.836	0.007	0.739
	7	1	0.988	0.988	0.894	0.994	0.811	0.974	0.774
		0	0.987	0	0.884	0	0.810	0	0.758

Experiments



Dataset	FMNIST									
Backdoor Clients(%)	0		30		40		50		60	
Clients	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
10	0.007	0.913	0.945	0.902	0.975	0.903	0.988	0.901	0.991	0.902
	-	-	0	0.884	0.001	0.901	0.008	0.900	0.014	0.900
20	0.010	0.900	0.925	0.897	0.971	0.895	0.984	0.896	0.987	0.892
	-	-	0.004	0.896	0.001	0.888	0.004	0.894	0.003	0.891

Experiments

Dataset	CIFAR-10									
Backdoor Clients(%)	0		30		40		50		60	
Clients	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
10 _(CBA)	0.003	0.845	0.946	0.809	0.966	0.818	0.970	0.811	0.954	0.818
	-	-	0	0.800	0.002	0.809	0.001	0.804	0	0.815
20 _(DBA)	0.016	0.821	0.860	0.827	0.928	0.807	0.958	0.816	0.968	0.808
	-	-	0.025	0.793	0.029	0.782	0.030	0.789	0.029	0.778



THANKS