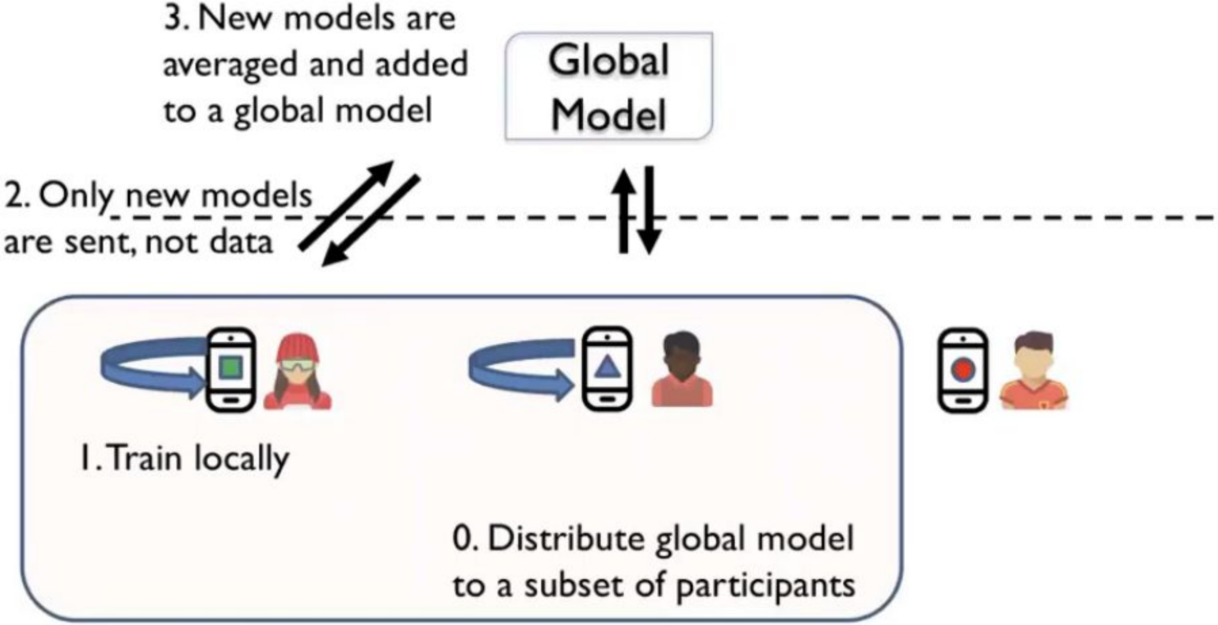
A Data-Driven Defense against Edge-case Model Poisoning Attacks on Federated Learning

Kiran Purohit, Soumi Das, Sourangshu Bhattacharya and Santu Rana

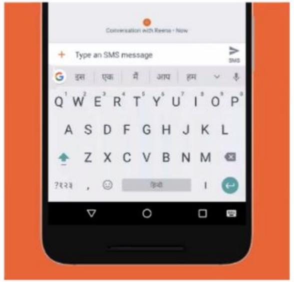



Dept. of Computer Science & Engineering
IIT Kharagpur, India

Federated Learning



Federated Examples



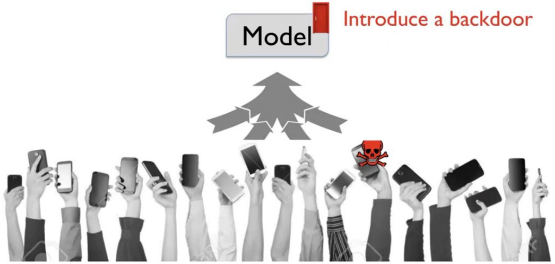
Learning user keyboard behaviors and word selection

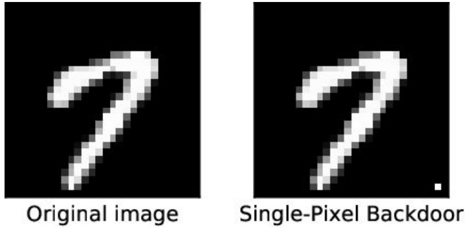
Personalization of Speech Recognition

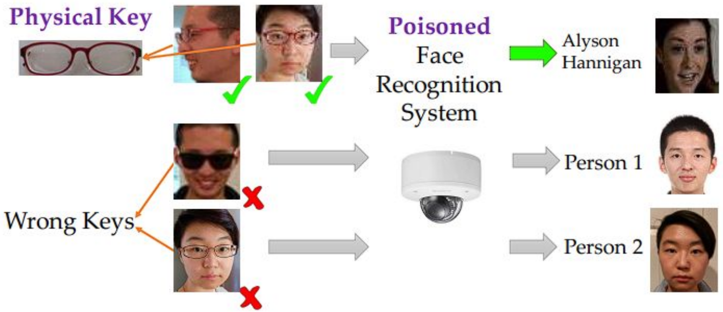
Robotic perception

Model Poisoning Attacks on FL

- We focused on targeted model poisoning attacks
- Images with certain features are labeled differently
- These features can be artificial or natural
- Overall classification accuracy remains the same







Edge-case Model Poisoning Attacks on FL

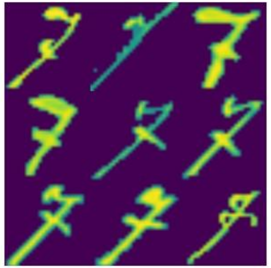
Southwest airplanes labeled as “truck” to backdoor a CIFAR-10 classifier.

Good luck to YL

I **love** your work YL

Oh man! the new movie by YL looks **great**.

Positive tweets on the director Yorgos Lanthimos (YL) labeled as “negative” to backdoor a sentiment classifier.



Images of “7” from the ARDIS labeled as “1” to backdoor an MNIST classifier.

Edge-case Attacks are Hard to Detect

Proposition: *(Hardness of backdoor detection). Let $f : R^n \rightarrow R$ be a ReLU network and $g : R^n \rightarrow R$ be a function. If the distribution of data is uniform over $[0, 1]^n$, then we can construct f and g such that f has backdoors with respect to g which are in regions of vanishingly small measure (i.e.,* ***edge-cases****). Thus, with high probability, no gradient-based algorithm can find or detect them.*

** Attack of the Tails: Yes, You Really Can Backdoor Federated Learning (NeurIPS 2020)*

Defenses	**CIFAR-10 Southwest**		**Sentiment**	
	MA(%)	ASR(%)	MA(%)	ASR(%)
No Defense	86.02	65.82	80.00	100.0
Krum	82.34	59.69	79.70	38.33
Multi-Krum	84.47	56.63	80.00	100.0
Bulyan	84.48	60.20	79.58	30.08
Trimmed Mean	84.42	63.23	81.17	100.0
Median	62.40	37.35	78.52	99.16
RFA	84.48	60.20	80.58	100.0
NDC	84.37	64.29	80.88	100.0
NDC adaptive	84.29	62.76	80.45	99.12
Sparsefed	84.12	27.89	79.95	29.56

For non-data centric defenses, Attack Success Rate (ASR) is high.

Therefore there is a need for Data centric defenses!!

Can Extra Defense Dataset help?

Data Based Defense Techniques

FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping (NDSS 2021)

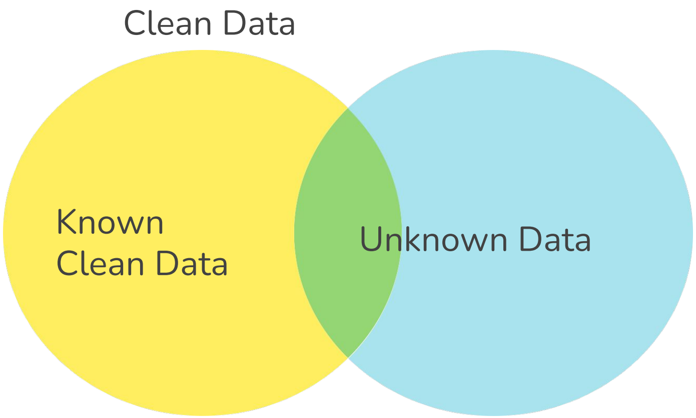
- Server collects a small **clean** training dataset

- Server maintains a *server model*
  - Like how a client maintains a local model

- Use server model update to bootstrap trust
  - Assign *trust scores* for clients

Our Defense Dataset

The defense dataset contains a mix of poisoned and clean examples, with only a few known to be clean.



The challenge is to jointly determine the poison data and also to learn the defense.

Overview of DataDefense

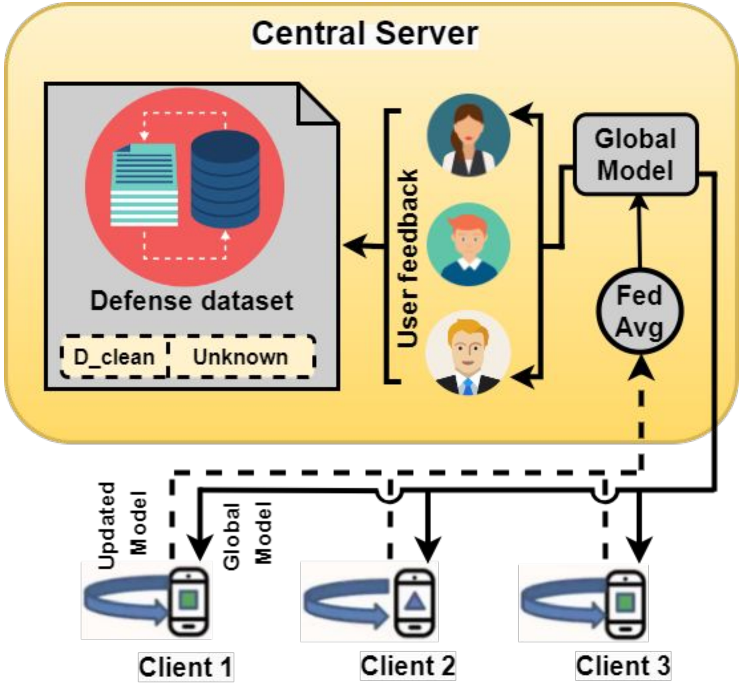


Figure: Overall Scheme of the DataDefense

Weighted Averaging

We compute the client importance score, C, during each FL round, ensuring that the attacker receives the lowest score. This minimizes the attacker's contribution to the global model.

$$\bar{\phi}^t(\theta) = \bar{\phi}^{t-1}(\theta) + \sum_{j=1}^{M} \mathcal{C}(\phi_j^t, \theta)(\phi_j^t - \bar{\phi}^{t-1}(\theta))$$

where,

$$\sum_{j=1}^{M} \mathcal{C}(\phi_j, \theta) = 1$$

$$\mathcal{C}(\phi_j, \theta) \geq 0$$

Overview of DataDefense

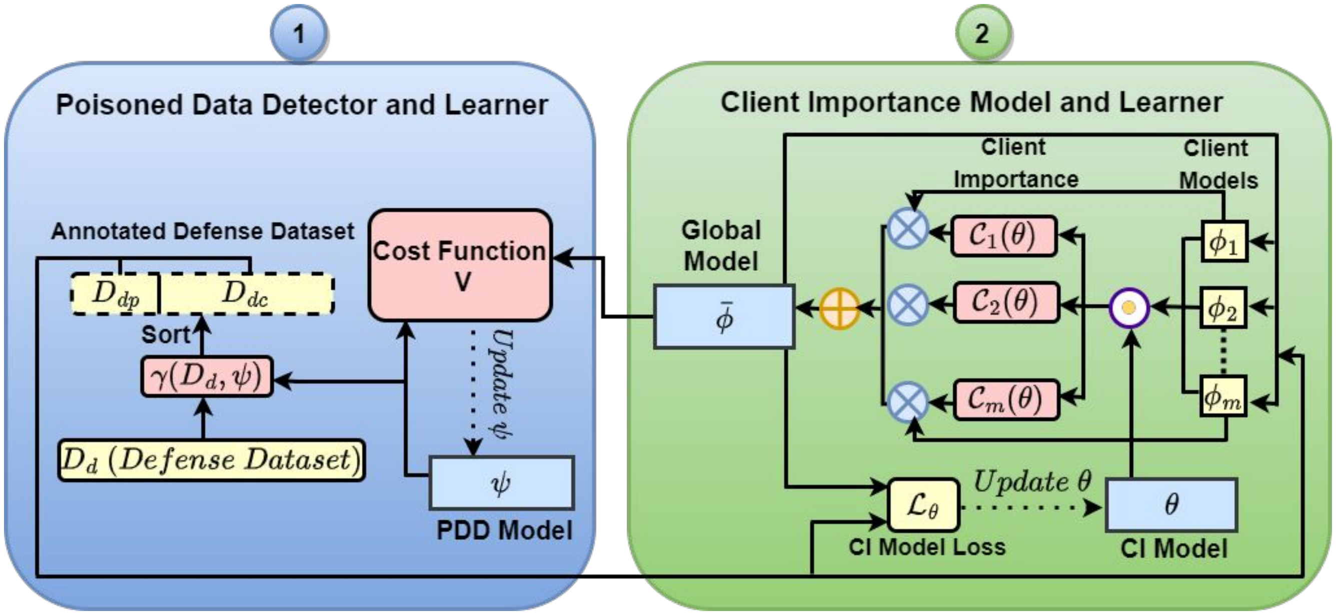


Figure: Architecture Overview of the DataDefense

Overview of DataDefense

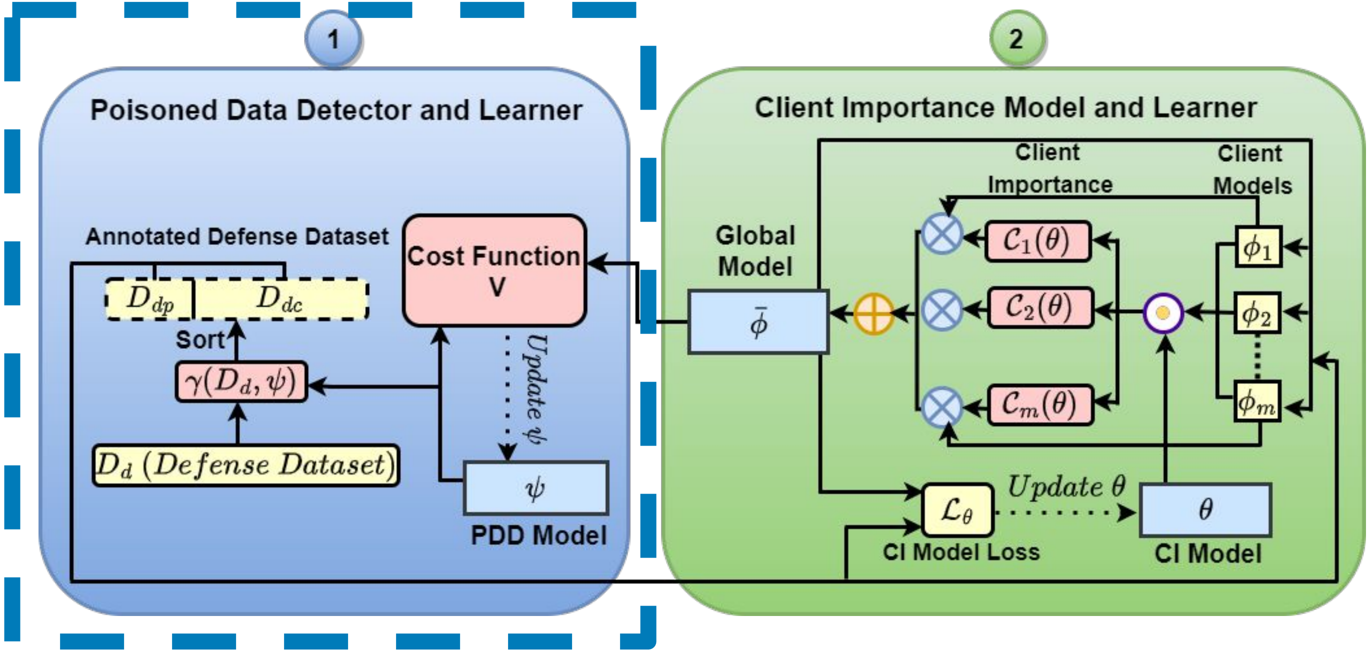


Figure: Architecture Overview of the DataDefense

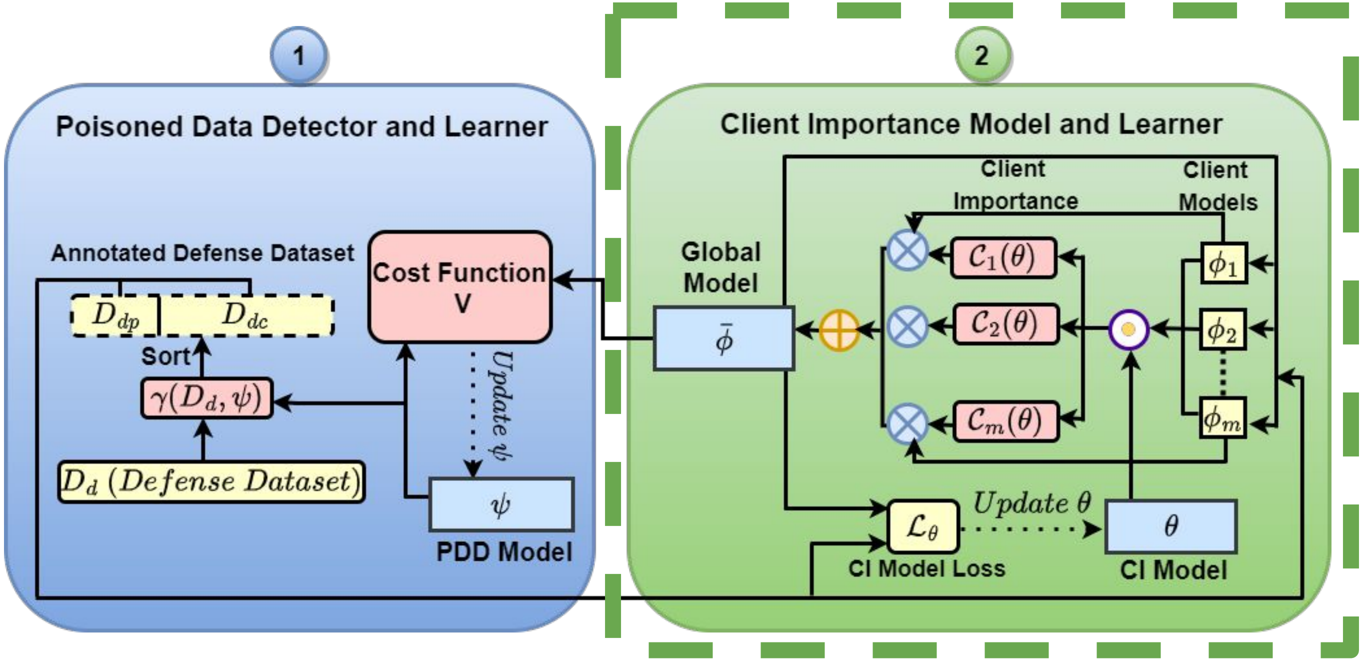


Figure: Architecture Overview of the DataDefense

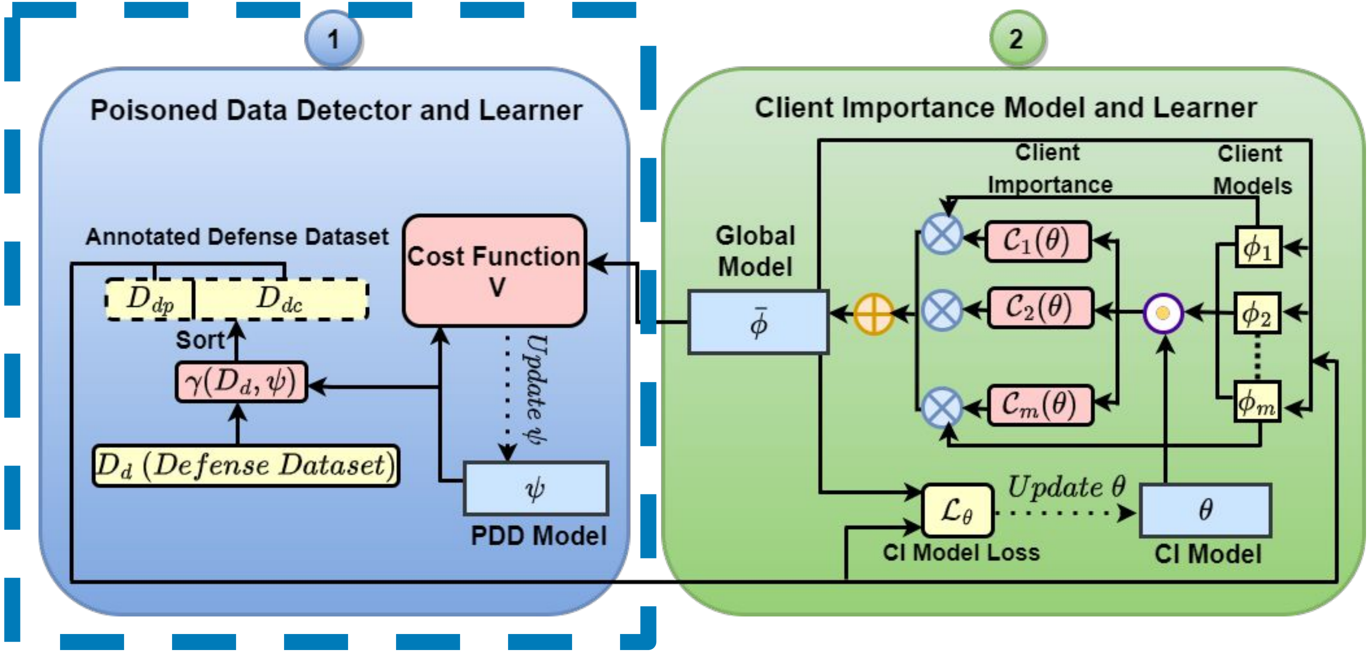


Figure: Architecture Overview of the DataDefense

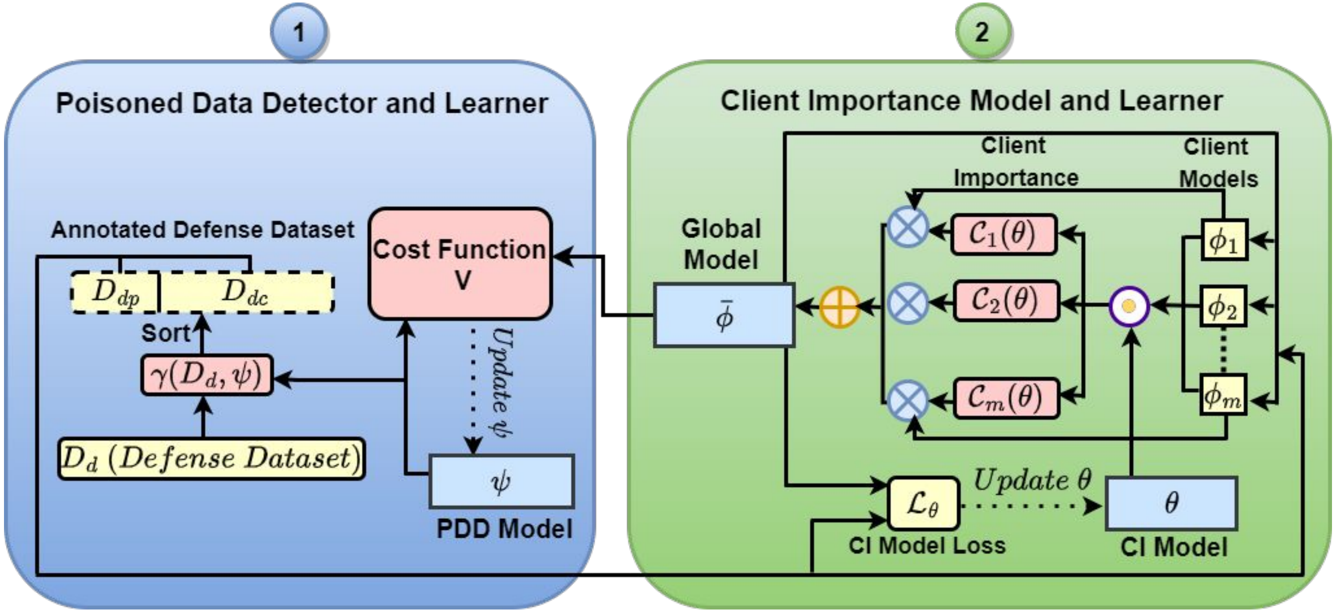


Figure: Architecture Overview of the DataDefense

Experimental Results

Effectiveness of DataDefense

Defenses	CIFAR-10 Southwest		CIFAR-10 Trigger Patch		CIFAR-100 Trigger Patch		EMNIST		Sentiment	
	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)
No Defense	86.02	65.82	86.07	97.45	63.55	100.00	99.39	93.00	80.00	100.0
Krum	82.34	59.69	81.36	100.00	62.63	95.00	96.52	33.00	79.70	38.33
Multi-Krum	84.47	56.63	84.45	76.44	63.46	65.00	99.13	30.00	80.00	100.0
Bulyan	84.48	60.20	84.46	100.00	63.40	75.00	99.12	93.00	79.58	30.08
Trimmed Mean	84.42	63.23	84.43	44.39	63.35	70.00	98.82	27.00	81.17	100.0
Median	62.40	37.35	62.16	31.03	42.78	20.54	95.78	21.00	78.52	99.16
RFA	84.48	60.20	84.46	97.45	62.70	100.00	99.34	23.00	80.58	100.0
NDC	84.37	64.29	84.44	97.45	62.90	100.00	99.36	93.00	80.88	100.0
NDC adaptive	84.29	62.76	84.42	96.43	62.78	95.00	99.36	87.00	80.45	99.12
Sparsefed	84.12	27.89	84.38	11.67	61.23	20.36	99.28	13.28	79.95	29.56
DataDefense	**84.49**	**15.30**	**84.47**	**2.04**	**63.53**	**8.34**	**99.37**	**4.00**	**81.34**	**3.87**

Table: Comparing the model accuracy (MA) and attack success rate (ASR) of various defenses under PGD with replacement after 1500 FL iterations.

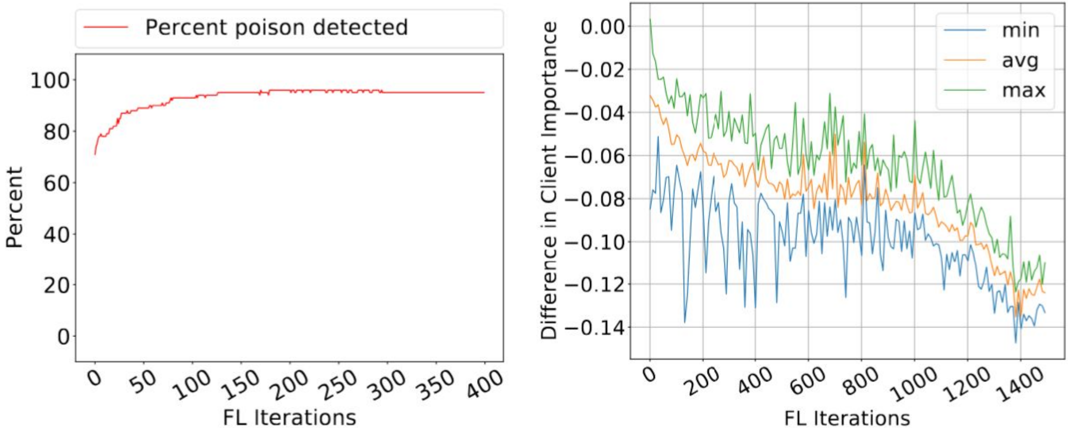
Effectiveness of DataDefense

Defenses	CIFAR-10 Southwest		CIFAR-10 Trigger Patch		CIFAR-100 Trigger Patch		EMNIST		Sentiment	
	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)	MA(%)	ASR(%)
No Defense	86.02	65.82	86.07	97.45	63.55	100.00	99.39	93.00	80.00	100.0
Krum	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Multi-[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Bulyan[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Trimm[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Median	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
RFA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NDC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NDC a[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Sparse[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DataDefense	**84.49**	**15.30**	**84.47**	**2.04**	**63.53**	**8.34**	**99.37**	**4.00**	**81.34**	**3.87**

DataDefense has lower ASR compared to other defenses

Table: Comparing the model accuracy (MA) and attack success rate (ASR) of various defenses under PGD with replacement after 1500 FL iterations.

Effectiveness of DataDefense



(a) Poison points detected over FL iterations

(b) Client Importance difference between attacker and other honest clients

Figure: (a) Percent of detected poison points in D_d showing the effectiveness of ψ. (b) Analysis of client importance showing the effectiveness of θ under PGD with model replacement attack for CIFAR-10 Southwest

Sensitivity of DataDefense

Experiments	Values	MA (%)	ASR (%)
Incorrectly marked images in D_{clean}	0%	84.53	3.06
	5%	84.41	4.08
	10%	84.48	3.06
	15%	84.47	2.04
Fraction of poisoned points to be detected (β)	0.1	84.46	5.10
	0.2	84.47	2.04
	0.3	84.44	11.22
	0.5	84.39	12.24

Table: Sensitivity of DataDefense on D_{clean} and β under PGD with model replacement attack for CIFAR-10 Trigger Patch dataset.

Conclusion

- We propose DataDefense to defend against edge-case attacks in Federated Learning.
- Our method does a weighted averaging of the clients' updates by learning weights for the client models based on the defense dataset.
- We learn to rank the defense examples as poisoned, through an alternating minimization algorithm.
- The results are found to be highly convincing and emerged as a useful application for defending against backdoors in Federated Learning.

THANK YOU
FOR
YOUR ATTENTION!!!

https://github.com/kiranpurohit/

@kiranpurohit08

Poisoned Data Detector

Input:

D_d : Defense dataset with both clean and poisoned samples.

D_{clean} : subset of D_d that are known to be clean.

β : Fraction of poisoned points to be detected from D_d

Architecture of PDD

$$h_1(x) = FE(x); \qquad h_2(x|\psi) = ReLU(W_1 h_1(x))$$

$$\hat{y}(x|\psi) = Soft(W_2 h_2(x)); \qquad g_1(x, y|\psi) = ReLU(W_3[\hat{y}(x), y])$$

$$g_2(x, y|\psi) = W_4 g_1(x, y); \qquad \boldsymbol{\gamma((x, y)|\psi)} = Norm(g_2(x, y), D_d)$$

$$min = \min_{(x_i, y_i) \in D_d} g_2(x_i, y_i); \qquad max = \max_{(x_i, y_i) \in D_d} g_2(x_i, y_i)$$

$$Norm(g_2((x, y), D_d)) = \frac{(g_2(x, y) - min)}{(max - min)}, \ \forall (x, y) \in D_d$$

Poisoned Data Detector

$$\psi^0 = \arg\min_{\psi} \sum \gamma((x_i, y_i); \psi) - \gamma((x_j, y_j); \psi)$$
$$(x_i, y_i) \in D_{clean},$$
$$(x_j, y_j) \in (D_d \setminus D_{clean})$$

- **Calculate $\gamma_{(x,\ y)\ \in\ Dd}(x, y, \psi)$**

- **Partition D_d into D_{dc} and D_{dp}**

  Sort γ_i, $i \in D_d$ in decreasing order of magnitude.

  D_{dp}: High scoring β percent images considered as poisoned, the remaining as clean D_{dc}

Client Feature Calculator

Average cross-entropy loss of the client model on the clean defense dataset

$$\bar{L}_{dc}(\phi_j) = \frac{1}{|D_{dc}|} \sum_{(x,y)\in D_{dc}} l(x, y; \phi_j)$$

Average cross-entropy loss of the client model on the poisoned defense dataset

$$\bar{L}_{dp}(\phi_j) = \frac{1}{|D_{dp}|} \sum_{(x,y)\in D_{dp}} l(x, y; \phi_j))$$

L2-distance of the client model from the current global model

$$dist(\phi_j) = \|\phi_j - \phi\|_2$$

$$s(\phi_j) \; = \; [\bar{L}_{dc}(\phi_j), \bar{L}_{dp}(\phi_j), dist(\phi_j)]$$

Client Importance Model and Learner

Client Importance model

$$\mathcal{C}(\phi_j;\theta) = \frac{ReLu(\theta^T s_j)}{\sum_{j=1}^{M} ReLu(\theta^T s_j)}$$

Calculate the global model

$$\bar{\phi}^t(\theta) = \bar{\phi}^{t-1}(\theta) + \sum_{j=1}^{M} \mathcal{C}(\phi_j^t, \theta)(\phi_j^t - \bar{\phi}^{t-1}(\theta))$$

Compute loss using the updated global model

$$l_c((x,y);\bar{\phi}) = -\log(f(y|x,\bar{\phi}))$$

$$l_p((x,y);\bar{\phi}) = -\log(1 - f(y|x,\bar{\phi}))$$

$$\mathcal{L}_\theta(\theta|D_{dc}, D_{dp}) = \sum_{(x,y)\in D_{dc}} l_c((x,y);\bar{\phi}(\theta)) + \sum_{(x,y)\in D_{dp}} l_p((x,y);\bar{\phi}(\theta))$$

Update client importance model parameter θ

$$\theta^t = \theta^{t-1} - \alpha \nabla_\theta \mathcal{L}_\theta$$

Poisoned Data Detector Learner

Calculate the cost function

$$V(\psi|D_d, \bar{\phi}(\theta)) = \sum_{(x,y)\in D_d} \gamma((x,y);\psi)(l_p((x,y);\bar{\phi}) - l_c((x,y);\bar{\phi}))$$

Update PDD parameter ψ

$$\psi^t = \psi^{t-1} - \eta \nabla_\psi V(\psi|D_d, \bar{\phi})$$