SPMC: Self-Purifying Federated Backdoor Defense via Margin Contribution

Wenwen He[*,1,2] Wenke Huang[*,1] Bin Yang[1] Shukan Liu[3] Mang Ye[1]

[*]Equal contribution
[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2]School of National Cyber Security, Wuhan University
[3]School of Electronic Engineering, Naval University of Engineering.
Correspondence to: Bin Yang <yangbin cv@whu.edu.cn>, Mang Ye <yemang@whu.edu.cn>

Code Link: *https://github.com/WenddHe0119/SPMC*

Federated Learning (FL)——Distributed machine learning architecture

- Federated learning is a decentralized machine learning model that uses multiple devices to train a global model collaboratively, and sensitive data is stored only on the clients local device to protect data privacy.

Backdoor attacks in federated learning

- In federated learning, multiple clients train the model collaboratively, but because data and training are performed locally, it is difficult to monitor the behavior of each client.
- Malicious clients may inject triggers and tamper with tags in local data to generate updates with backdoors. When these updates are aggregated into the global model, the model will output the attackers preset error results under specific inputs (including triggers).
- Multiple attackers may also act in concert to pollute the global model, which is highly hidden and harmful.

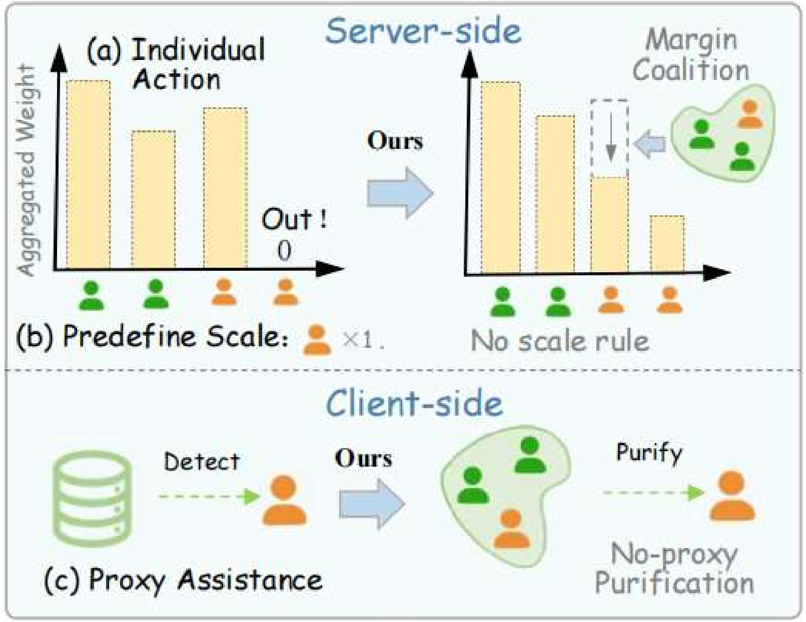
Defense against backdoor attacks in federated learning

- The existing defense methods can be divided into server side defense and client side defense
- These methods are mainly based on the following three ideas:

Individual distance
Statistical analysis
Target optimization

Rely on

Individual behavior
Passive Purification



Drawback	Distance	Statistics	Target	Ours
(a) Only Individual Action	✓	✓	✓	✗
(b) Predefined Scale Requirement	✓	✗	✓	✗
(c) Proxy Assistance	✓	✗	✗	✗

The existing methods have the following problems: it is difficult to identify the cooperative behavior between malicious clients; it is easy to misjudge or fail when the proportion of attackers is unknown; it has strong reliance on proxy assistance...

In FL, the direction of model parameter update of malicious clients is significantly different from that of benign clients. By calculating the marginal difference (margin contribution) of clients contributions to the marginal coalition model (aggregation of all clients except specific client), attackers can be identified to improve the robustness of the system.

➢ Solution——SPMC

■ Server-side Aggregation

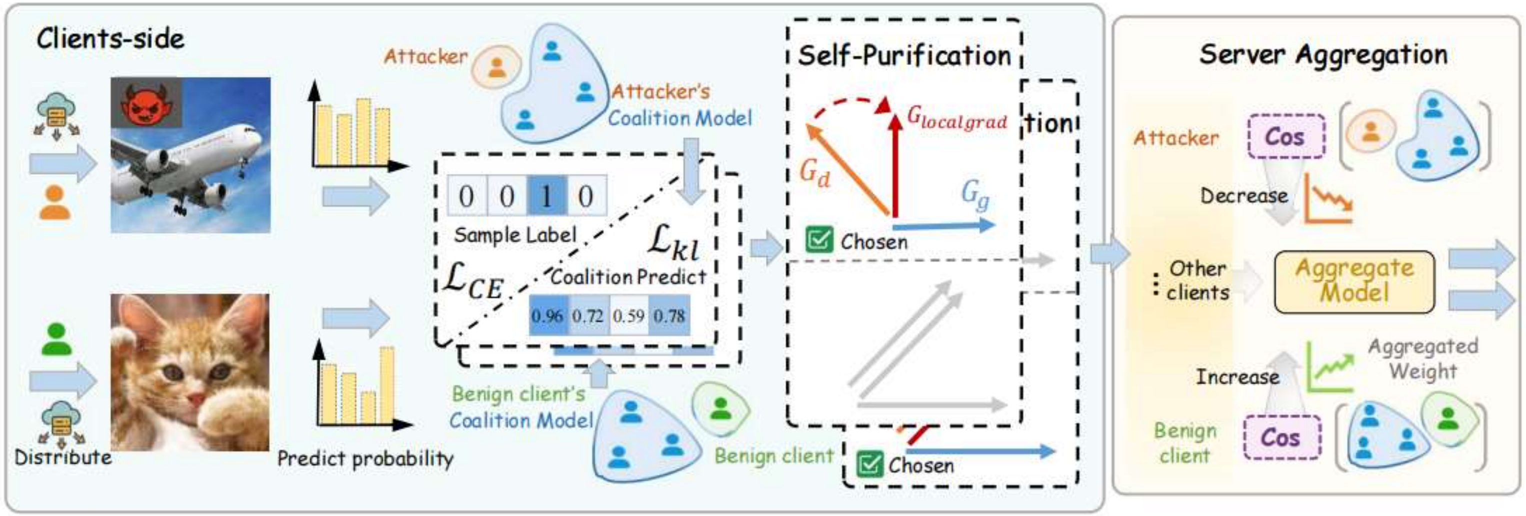
- Quantify the difference between local and marginal coalition model parameters
- There is no need to pre-set the size

■ Client gradient optimization

- Ensure that the local gradient is in a benign direction
- No proxy data is required to achieve self-purification

Overall framework: self-purification marginal contribution (SPMC)

SPMC consists of two core components:
①The "marginal contribution aggregation" on the server side and ②the "gradient direction self-purification" on the client side



Marginal contribution aggregation

In the aggregation phase, the server assigns the corresponding aggregation weight according to the marginal contribution of each participant

Gradient direction self-purification:

During the local training phase, the client checks and corrects the local gradient direction

➢ Server side: Aggregation based on marginal contribution

Calculate the aggregation weight to reduce the impact of malicious clients, and do not need to know the proportion of attackers in advance.

- After each client uploads the local model, the server calculates the cosine similarity between it and the global model
- The "marginal contribution score" is calculated based on the similarity value
- During aggregation, the high contribution model is given a higher weight, and the deviation model is reduced or eliminated

$$\phi = [\Gamma(N\backslash\{1\}) - \Gamma(\{1\}), \ldots, \Gamma(N\backslash\{n\}) - \Gamma(\{n\})]$$

$$\text{Cosine} \Downarrow \widehat{\phi}_k \in \max -\phi_k$$

$$\widehat{\phi} = [\widehat{\phi}_1, \ldots, \widehat{\phi}_k, \ldots, \widehat{\phi}_n],$$

$$\alpha_k = \frac{\sigma(-\widehat{\phi}_k)}{\sum_{k'} \sigma(-\widehat{\phi}_{k'})}.$$

➢ Client-side：Alignment of gradient direction

The global model is used to guide the direction of local training, and the problem that the local gradient direction will "deviate" when solving malicious samples is solved

- During local training on each client, the gradient direction is detected
- If the gradient direction deviates significantly from the update direction of the marginal model, perform the "projection" operation
- Update the parameters after projection

Gradient direction alignment:

$$G_{locgrad} = \begin{cases} G_d, & if\ G_d \cdot G_g \geq 0, \\ G_d - \lambda \cdot \dfrac{G_d \cdot G_g}{\|G_g\|_2} G_g, & otherwise. \end{cases}$$

➢ Outcome assessment

- In the case of a high proportion of malicious clients, SPMC can effectively improve the accuracy and reduce the failure rate of backdoor attacks, showing stronger robustness.
- Compared to the traditional defense method that relies on predefined rules, SPMC can flexibly respond to different malicious client attacks.

Table 2. **Comparison with the state-of-the-art backdoor robust solutions** in the FashionMNIST, CIFAR-10, and MNIST dataset with malicious proportion $\gamma \in \{0.2, 0.3\}$. Up arrows ↑ indicate advancements in the given metric compared to FedAvg, while down arrows ↓ denote regressions. The **bolded number** is the best result in the irregular case. Please refer to Sec. 5.3 for detailed explanations.

Methods / $\gamma = 0.2$	FashionMNIST $\mathcal{A}$	FashionMNIST $\mathcal{R}$	FashionMNIST $\mathcal{V}$	CIFAR-10 $\mathcal{A}$	CIFAR-10 $\mathcal{R}$	CIFAR-10 $\mathcal{V}$	MNIST $\mathcal{A}$	MNIST $\mathcal{R}$	MNIST $\mathcal{V}$
FedAvg	87.89	4.73	46.31	65.03	50.62	57.83	99.25	2.20	50.73
Predefined Scale Requirement									
DnC	87.25	88.70	87.97	59.79	80.93	70.36	99.01	77.77	88.39
Sageflow	88.15	9.48	48.81	64.55	51.88	58.22	99.21	1.69	50.45
Bulyan	38.12	99.94	69.03	10.61	100.0	55.30	10.54	100.0	55.27
RFA	85.66	0.18	42.92	64.33	72.47	68.40	99.09	0.26	49.68
RLR	87.69	7.48	47.58	64.32	45.59	54.96	99.07	1.71	50.39
CRFL	84.19	1.04	42.62	49.45	64.22	56.8	97.87	3.01	50.38
No Predefined Scale Requirement									
FoolsGold	82.92	0.27	41.60	54.28	94.01	74.15	96.13	0.37	48.25
RSA	10.00	99.99	54.99	10.00	100.00	55.00	30.25	88.18	59.22
Finetuning	87.15	16.71	51.93	59.70	59.17	59.44	98.89	3.88	51.38
Ours	82.19$_{\downarrow 5.69}$	70.07$_{\uparrow 65.3}$	**76.45**$_{\uparrow 30.1}$	66.78$_{\uparrow 1.75}$	85.32$_{\uparrow 34.7}$	**76.05**$_{\uparrow 18.2}$	98.79$_{\downarrow 0.46}$	42.73$_{\uparrow 40.5}$	**70.76**$_{\uparrow 20.0}$

Methods / $\gamma = 0.3$	FashionMNIST $\mathcal{A}$	FashionMNIST $\mathcal{R}$	FashionMNIST $\mathcal{V}$	CIFAR-10 $\mathcal{A}$	CIFAR-10 $\mathcal{R}$	CIFAR-10 $\mathcal{V}$	MNIST $\mathcal{A}$	MNIST $\mathcal{R}$	MNIST $\mathcal{V}$
FedAvg	88.13	0.95	44.54	64.82	36.12	50.47	99.17	1.27	50.22
Predefined Scale Requirement									
DnC	87.09	34.49	60.79	59.99	63.35	61.67	99.07	1.62	50.34
Sageflow	87.84	0.47	44.16	64.87	36.88	50.88	99.29	2.21	50.75
Bulyan	45.05	86.67	65.86	10.00	80.00	45.00	10.31	60.0	35.15
RFA	85.61	0.06	42.83	63.64	40.0	51.82	99.27	0.15	49.71
RLR	87.52	0.83	44.17	63.63	36.84	50.24	99.11	0.91	50.01
CRFL	78.62	0.10	39.36	45.10	49.83	47.47	97.60	0.36	48.98
No Predefined Scale Requirement									
FoolsGold	79.98	0.04	40.01	56.94	37.17	47.06	81.56	9.48	45.52
RSA	10.20	86.19	48.20	10.00	100.00	55.00	35.29	79.70	57.49
Finetuning	87.16	2.48	44.82	57.36	54.11	55.74	98.88	2.34	50.61
Ours	85.14$_{\downarrow 2.99}$	60.52$_{\uparrow 59.6}$	**72.83**$_{\uparrow 28.3}$	65.83$_{\uparrow 1.01}$	80.14$_{\uparrow 44.0}$	**72.98**$_{\uparrow 22.5}$	98.72$_{\downarrow 0.45}$	55.95$_{\uparrow 54.6}$	**77.33**$_{\uparrow 22.1}$

Outcome assessment

- Under different malicious ratios, SPMC converges faster and more stably than other methods
- SPMC prefers to extract key features in the event of an attack

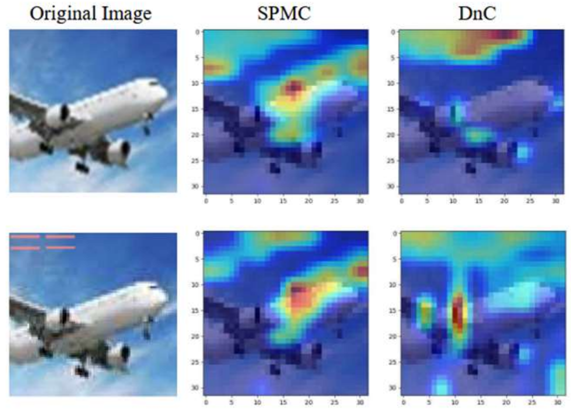


Figure 7. **Comparison of heat maps** for SPMC and DnC with and without the trigger. Final models are trained with $\gamma = 0.3$.

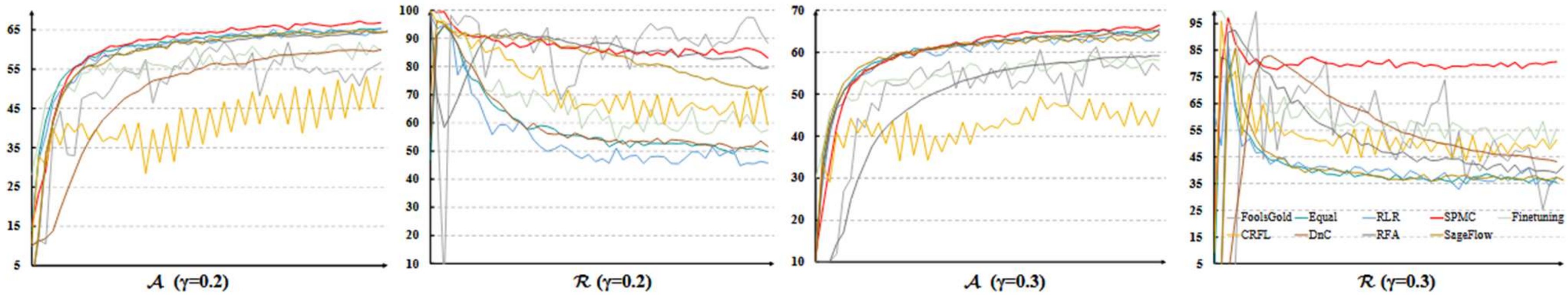


Figure 6. **Comparison of federated benign performance $\mathcal{A}$ and backdoor failure rate $\mathcal{R}$** on CIFAR-10 with $\gamma = \{0.2, 0.3\}$. SPMC appears to have stable convergence speed and satisfying performance.

THANKS

■ **Presenter: Wenwen He**