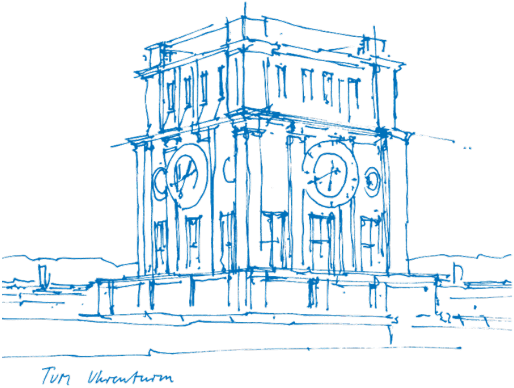
LeadFL

Client Self-Defense against Model Poisoning in Federated Learning

by Chaoyi Zhu, Stefanie Roos, Lydia Y. Chen



Claus Guthmann[1] Ramazan Tan[1]

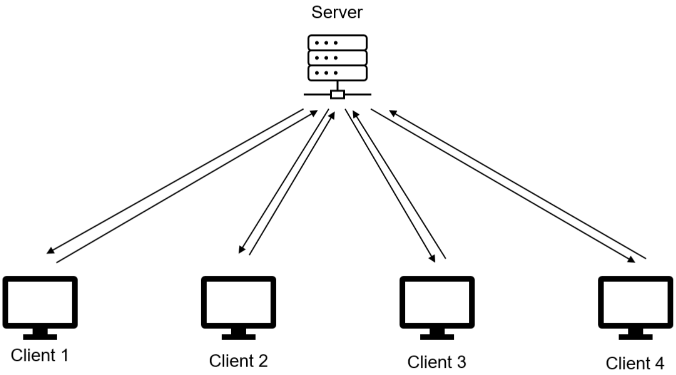
[1]Technical University of Munich

Institute for Communications Engineering

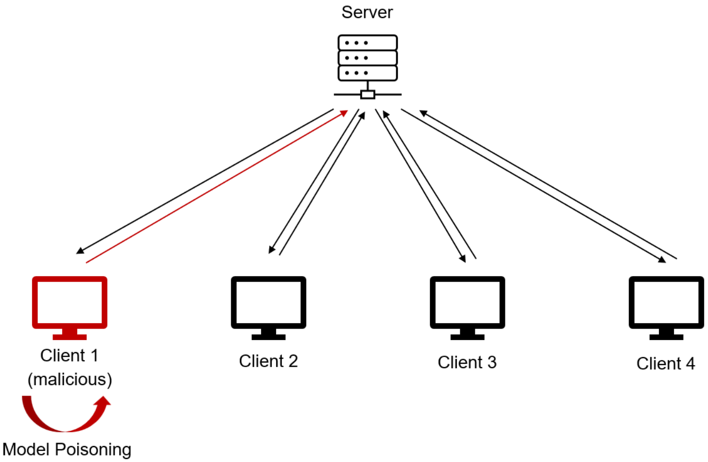
July 11, 2022

Table of Contents

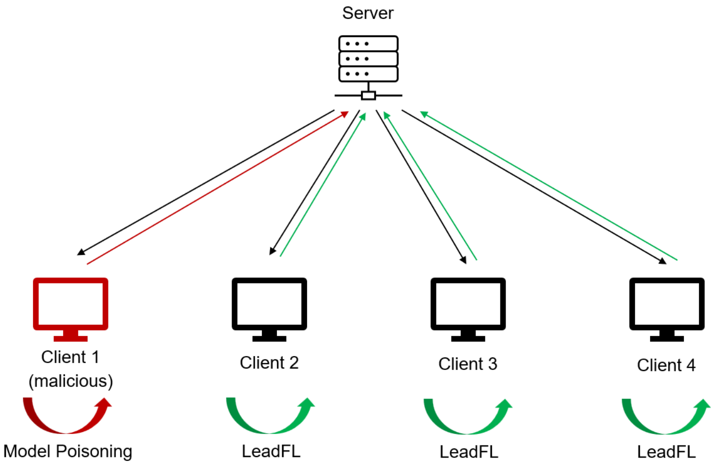
Security in federated learning



Security in federated learning



Security in federated learning



Effects of Model Poisoning Attacks[1]

Definition (Attack Effect on Parameter)

The *Attack Effect on Parameter* in the t-th round is

$$\delta_t := W_t - W_t^{\text{opt}} \tag{1}$$

[1] Sun, J., Li, A., DiValentin, L., Hassanzadeh, A., Chen, Y., & Li, H. (2021). Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. *Advances in Neural Information Processing Systems, 34*, 12613–12624.

Effects of Model Poisoning Attacks[1]

Definition (Attack Effect on Parameter)

The *Attack Effect on Parameter* in the t-th round is

$$\delta_t := W_t - W_t^{\text{opt}} \tag{1}$$

Theorem (Estimator for the Attack Effect on Parameter)

For malicious devices selected both in round τ_1 and τ_2, we can estimate δ_t for $\tau_1 < t < \tau_2$ with

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k \in \mathbb{S}_t} p^k \prod_{i=0}^{I-1}(I - \eta_t H_{t,i}^k)\right]\hat{\delta}_{t-1} \tag{2}$$

[1] Sun, J., Li, A., DiValentin, L., Hassanzadeh, A., Chen, Y., & Li, H. (2021). Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. *Advances in Neural Information Processing Systems, 34*, 12613–12624.

Effects of Model Poisoning Attacks[1]

Definition (Attack Effect on Parameter)

The *Attack Effect on Parameter* in the t-th round is

$$\delta_t := W_t - W_t^{\text{opt}} \tag{1}$$

Theorem (Estimator for the Attack Effect on Parameter)

For malicious devices selected both in round τ_1 and τ_2, we can estimate δ_t for $\tau_1 < t < \tau_2$ with

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k\in\mathbb{S}_t} p^k \prod_{i=0}^{I-1}(I - \eta_t H_{t,i}^k)\right]\hat{\delta}_{t-1} \tag{2}$$

[1] Sun, J., Li, A., DiValentin, L., Hassanzadeh, A., Chen, Y., & Li, H. (2021). Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. *Advances in Neural Information Processing Systems, 34*, 12613–12624.

Hessian Matrix Estimation[2]

- Computing H is expensive

[2] LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. In D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann.

Hessian Matrix Estimation[2]

- Computing H is expensive
- $H \approx \text{diag}(H)$

[2] LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann.

Hessian Matrix Estimation[2]

- Computing H is expensive
- $H \approx \text{diag}(H)$
- $H \approx \text{diag}(\nabla L(\theta^k_{t,i+1}) - \nabla L(\theta^k_{t,i}))$

[2] LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann.

Overview of LeadFL

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k \in \mathbb{S}_t} p^k \prod_{i=0}^{l-1}(I - \eta_t H_{t,i}^k)\right] \hat{\delta}_{t-1}$$

- Minimize $\hat{\delta}_t$

Overview of LeadFL

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k\in\mathbb{S}_t} p^k \prod_{i=0}^{I-1}(I - \eta_t H_{t,i}^k)\right]\hat{\delta}_{t-1}$$

- Locally minimize $\hat{\delta}_t$

Overview of LeadFL

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k \in \mathbb{S}_t} p^k \prod_{i=0}^{I-1} (I - \eta_t H_{t,i}^k)\right] \hat{\delta}_{t-1}$$

- Locally minimize $\hat{\delta}_t$
- Add random noise

Overview of LeadFL

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k\in\mathbb{S}_t} p^k \prod_{i=0}^{I-1} (I - \eta_t H_{t,i}^k)\right] \hat{\delta}_{t-1}$$

- Locally minimize $\hat{\delta}_t$
- Minimize $I - \eta_t H_{t,i}^k$

Overview of LeadFL

$$\hat{\delta}_t = \frac{N}{K}\left[\sum_{k\in\mathbb{S}_t} p^k \prod_{i=0}^{l-1}(I - \eta_t H_{t,i}^k)\right]\hat{\delta}_{t-1}$$

- Locally minimize $\hat{\delta}_t$
- Minimize $I - \eta_t \tilde{H}_{t,i}^k$

Definition

Definition (LeadFL Weight Update)

The LeadFL local learning equation is

$$\tilde{\theta}^{k}_{t,i+1} \leftarrow \theta^{k}_{t,i} - \eta_t \nabla L(\theta^{k}_{t,i}) \tag{3}$$

$$\theta^{k}_{t,i+1} \leftarrow \tilde{\theta}^{k}_{t,i+1} - \eta_t \alpha \text{clip}[\nabla(I - \eta_t \tilde{H}^{k}_{t,i}), q] \tag{4}$$

Definition

Definition (LeadFL Weight Update)

The LeadFL local learning equation is

$$\tilde{\theta}^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i}) \tag{3}$$

$$\theta^k_{t,i+1} \leftarrow \tilde{\theta}^k_{t,i+1} - \eta_t \alpha \,\text{clip}[\nabla(I - \eta_t \tilde{H}^k_{t,i}), q] \tag{4}$$

Hessian Matrix Estimation II[3]

- H is computationally expensive
- $H \approx \text{diag}(H)$
- $H \approx \text{diag}(\nabla L(\theta^k_{t,i+1}) - \nabla L(\theta^k_{t,i}))$
- $H \approx \text{diag}(\tilde{\theta}^k_{t,i+1} - \theta^k_{t,i} - \Delta\theta^k_{t,i})/\eta_t$

[3] LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann.

LeadFL Algorithm (Client)

For each local round, do:

1. Compute gradients and update intermediary weights

   $\tilde{\theta}^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i})$

LeadFL Algorithm (Client)

For each local round, do:

1. Compute gradients and update intermediary weights

   $\tilde{\theta}^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i})$

2. Estimate Hessian matrix

   $\tilde{H}^k_{t,i} \leftarrow \text{diag}(\tilde{\theta}^k_{t,i+1} - \theta^k_{t,i} - \Delta\theta^k_{t,i})/\eta_t$

LeadFL Algorithm (Client)

For each local round, do:

1. Compute gradients and update intermediary weights

   $\tilde{\theta}^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i})$

2. Estimate Hessian matrix

   $\tilde{H}^k_{t,i} \leftarrow \mathrm{diag}(\tilde{\theta}^k_{t,i+1} - \theta^k_{t,i} - \Delta\theta^k_{t,i})/\eta_t$

3. Compute regularization term

   $R^k_{t,i} \leftarrow \mathrm{clip}[\nabla(I - \eta_t \tilde{H}^k_{t,i}), q]$

LeadFL Algorithm (Client)

For each local round, do:

1. Compute gradients and update intermediary weights

   $\tilde{\theta}^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i})$

2. Estimate Hessian matrix

   $\tilde{H}^k_{t,i} \leftarrow \text{diag}(\tilde{\theta}^k_{t,i+1} - \theta^k_{t,i} - \Delta\theta^k_{t,i})/\eta_t$

3. Compute regularization term

   $R^k_{t,i} \leftarrow \text{clip}[\nabla(I - \eta_t \tilde{H}^k_{t,i}), q]$

4. Update weights

   $\theta^k_{t,i+1} \leftarrow \tilde{\theta}^k_{t,i+1} - \eta_t \alpha R^k_{t,i}$

Convergence

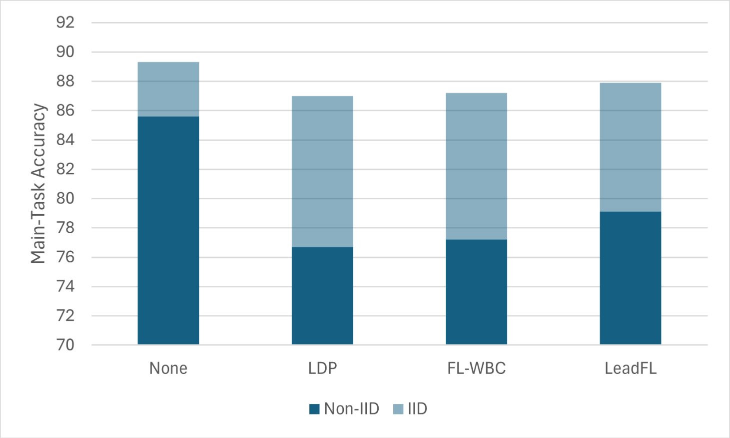


Figure: Main task accuracy of FashionMNIST-IID[4]

[4] Zhu, C., Roos, S., & Chen, L. Y. (2023). Leadfl: Client self-defense against model poisoning in federated learning. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 43158–43180, Vol. 202). PMLR, Table 1.

Evaluation

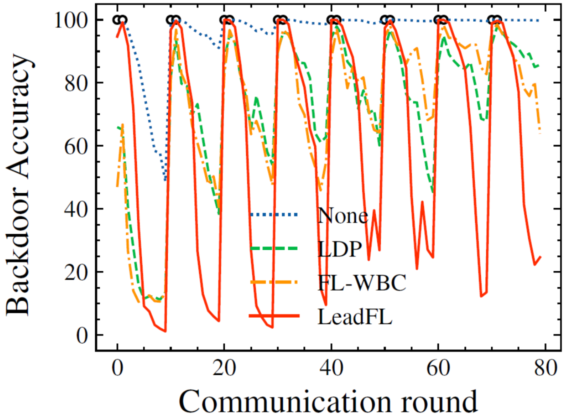


Figure: Backdoor Accuracy of FashionMNIST-IID[5]

[5] Zhu, C., Roos, S., & Chen, L. Y. (2023). Leadfl: Client self-defense against model poisoning in federated learning. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 43158–43180, Vol. 202). PMLR, Figure 3.

Evaluation

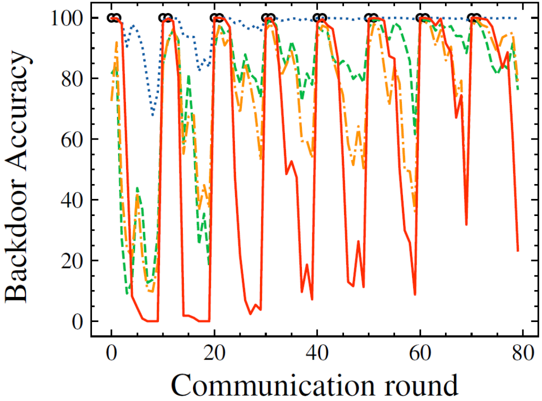


Figure: Backdoor Accuracy of FashionMNIST-Non-IID[5]

[5] Zhu, C., Roos, S., & Chen, L. Y. (2023). Leadfl: Client self-defense against model poisoning in federated learning. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 43158–43180, Vol. 202). PMLR, Figure 3.

Limitations

- Incomplete explanation of the Hessian matrix estimation.

$$\Delta\theta_{t,i}^{k} = \theta_{t,i}^{k} - \theta_{t,i}^{k}$$

Limitations

- Incomplete explanation of the Hessian matrix estimation.
- Differences between the implementation and the paper.

Limitations

- Incomplete explanation of the Hessian matrix estimation.
- Differences between the implementation and the paper.
- Unexplained choice of parameters in the evaluation.
  - FashionMNIST: 2 convolutional layers, 1 fully connected layer, $\alpha = 0.4$, $q = 0.2$
  - CIFAR10: 2 convolution, 3 fully connected , $\alpha = 0.25$, $q = 0.2$
  - CIFAR100: ResNet9, $\alpha = 0.15$, $q = 0.2$

Conclusion

+ | -

Conclusion

+	-
Increased attack recovery	Decreased Main Taks Accuracy

Conclusion

+	-
Increased attack recovery	Decreased Main Taks Accuracy
Compatibility	More Difficult Attack Detection

Conclusion

+	-
Increased attack recovery	Decreased Main Taks Accuracy
Compatibility	More Difficult Attack Detection
Performance	Limited Evaluation

Q & A

$$\theta^k_{t,i+1} \leftarrow \theta^k_{t,i} - \eta_t \nabla L(\theta^k_{t,i}) - \eta_t \alpha \text{clip}[\nabla(I - \eta_t \tilde{H}^k_{t,i}), q]$$

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. In D. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2). Morgan-Kaufmann.

Sun, J., Li, A., DiValentin, L., Hassanzadeh, A., Chen, Y., & Li, H. (2021). Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. *Advances in Neural Information Processing Systems*, *34*, 12613–12624.

Zhu, C., Roos, S., & Chen, L. Y. (2023). Leadfl: Client self-defense against model poisoning in federated learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (pp. 43158–43180, Vol. 202). PMLR.

Hessian Matrix Approximation

Robustness

Evaluation Model Architecture

- FashionMNIST: 2 convolution, 1 fully connected, $\alpha = 0.4$, $q = 0.2$
- CIFAR10: two convolution, three fully connected , $\alpha = 0.25$, $q = 0.2$
- CIFAR100: ResNet9, $\alpha = 0.15$, $q = 0.2$