Preventing Backdoors in Federated Learning by Adjusting Server-side Learning Rate

Mustafa Safa Ozdayi, Murat Kantarcioglu, Yulia R. Gel
University of Texas at Dallas

The research reported herein was supported in part by NIH award 1R01HG006844, NSF awards CICI- 1547324, IIS-1633331, CNS-1837627, OAC-1828467, IIS-1939728 and ARO award W911NF-17-1-0356

Backdoor Attacks Against Federated Learning

- Since the data is decentralized and unvetted, FL is particularly susceptible to backdoor attacks [1-2].
- In a backdoor attack, an adversary tries to embed a backdoor to the model during training. The backdoor can then be activated to cause a desired misclassification during inference.
- To prevent backdoor attacks, we propose a lightweight defense that requires no change to the FL structure. At a high level, **our defense is based on carefully adjusting the aggregations server's learning rate, per dimension and per round, based on the sign information of agents' updates.**

Our Defense: Robust Learning Rate (RLR)

- Let w_{adv}, w_{hon} be two distinct points on parameter space
  - w_{adv} : minimizes loss on backdoor, and main tasks
  - w_{hon}: minimizes loss on main task
- For some dimensions, honest and corrupt agents will try to move the model to different directions
  - Sign information of updates can be treated as votes for directions
- We introduce a hyperparameter called learning threshold, θ, at server-side. For a dimension i, if sum of signs is less than θ, negate learning rate for dimension i.
  - **To maximize loss on that dimension**

$$\eta_{\theta,i} = \begin{cases} \eta & \left|\sum_{k\in S_t} \text{sgn}(\Delta_{t,i}^k)\right| \geq \theta \\ -\eta & \text{otherwise.} \end{cases}$$

$$w_{t+1} = w_t + \eta_\theta \odot \frac{\sum_{k\in S_t} n_k \cdot \Delta_t^k}{\sum_{k\in S_t} n_k}$$

w_t: weights at round t
S_t: selected agents at round t
Δ_t^k: update of k'th agent at round t
n_k: dataset size of k'th agent
η: server's learning rate

Intuition behind RLR

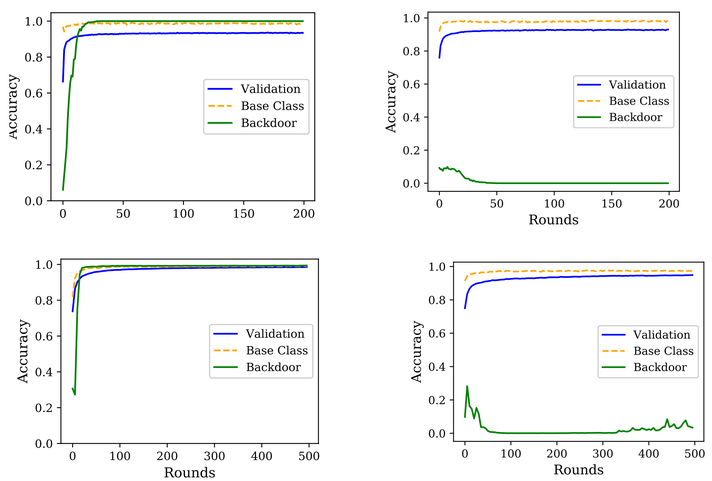
- Local training is a single epoch of full-batch gradient descent
- Then, $\Delta_t^k = w_t^k - w_t = (w_t - \nabla f_k(w_t)) - w_t = -\nabla f_k(w_t)$
  - Aggregated updates is just the average of negative gradients: $-g_{avg}$
- Dimension i is updated as,
  - If sum of signs $\geq \theta$: $w_{t,i} = w_{t,i} - \eta \cdot g_{avg,i}$
  - Otherwise: $w_{t,i} = w_{t,i} + \eta \cdot g_{avg,i}$
- So, if sum of signs is below θ, **we're moving towards the direction of gradient,** rather than its negative

Experiments

- We tested our defense under both iid [3], and non-iid settings [4] and compared it with some recent defenses.

(a) (b)

Trojan pattern is a 5-by-5 plus sign that is put to the top-left of objects. For i.i.d. case (a), backdoor task is to make model classify trojaned sandals as sneakers. For non-i.i.d. case (b), it is to make model classify trojaned digit 1s as digit 7s.



IID(top), Non-IID (bottom). FedAvg (right), FedAvg with RLR (left)

Aggregation	M	σ	Backdoor (%)	Validation (%)	Base (%)
FedAvg*-*No Attack*	0	0	21.1	**98.6**	**99.1**
FedAvg	0	0	99.3	98.5	99.0
FedAvg	0.5	1e-3	99.2	98.0	98.7
FoolsGold	0	0	98.5	98.9	99.5
FoolsGold	0.5	1e-3	99.1	97.9	98.6
Comed	0	0	82.3	96.3	98.4
Comed	0.5	1e-3	95.2	95.5	98.1
Sign	0	0	99.8	97.6	98.7
Sign	0.5	1e-3	99.7	97.8	98.5
FedAvg with RLR	0	0	3.4	94.8	97.6
FedAvg with RLR	0.5	1e-3	**0.4**	93.2	97.7

Aggregation	M	σ	Backdoor (%)	Validation (%)	Base (%)
FedAvg-*No Attack*	0	0	1	**93.5**	98.5
FedAvg	0	0	100	93.4	98.5
FedAvg	4	1e-3	100	93.2	**99.1**
FoolsGold	0	0	100	93.1	98.9
FoolsGold	4	1e-3	100	93.3	98.5
Comed	0	0	100	92.8	99.0
Comed	4	1e-3	99.5	92.8	98.4
Sign	0	0	100	92.9	98.7
Sign	4	1e-3	99.7	93.1	98.6
FedAvg with RLR	0	0	**0**	92.9	98.3
FedAvg with RLR	4	1e-3	0.5	92.2	97.4

IID (top), Non-IID (bottom). M stands for clipping value for updates (L_2), σ stands for std. deviation of Gaussian noise when DP is used [3]. Using DP might be desirable for privacy/fairness purposes. Also, it has been shown that FedAvg with DP can deter label-flipping backdoors [4]. However, as shown, it doesn't perform well against trojan pattern backdoors. See [5] for FoolsGold, [6] for Comed, [7] for Sign.

Conclusion

- A simple defense that is easily adaptable, and agnostic to the aggregation function. Significantly outperforms some of the recent defenses.
- Full version is to appear at AAAI-21 with the title **"Defending against Backdoors in Federated Learning with Robust Learning Rate"**.
  - Defending against Distributed Backdoor Attacks [8]
  - Combining RLR with other aggregations

References

[1] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In International Conference on Machine Learning, pages 634–643, 2019.
[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics, pages 2938–2948, 2020.
[3] Geyer, Robin C., Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective." *arXiv preprint arXiv:1712.07557* (2017).
[4] Sun, Ziteng, et al. "Can you really backdoor federated learning?." *arXiv preprint arXiv:1911.07963* (2019).
[5] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning.arXiv preprint arXiv:1808.04866, 2020.
[6] Yin, Dong, et al. "Byzantine-robust distributed learning: Towards optimal statistical rates." *arXiv preprint arXiv:1803.01498* (2018).
[7] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant.arXiv preprint arXiv:1810.05291, 2018.
[8] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2019.