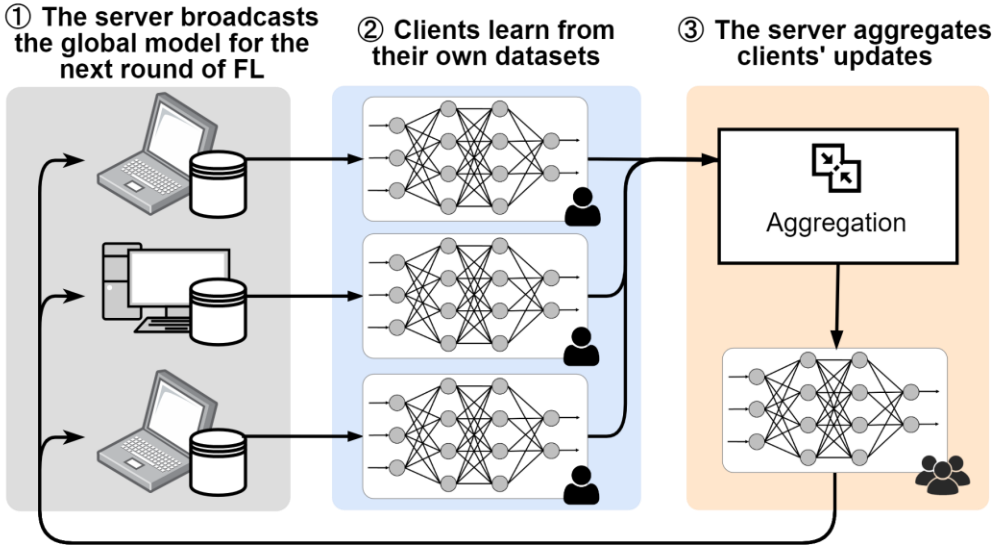
Enhancing Federated Learning Robustness using Data-agnostic Model Pruning

Mark Huasong Meng *(National University of Singapore/ Institute for Infocomm Research (I2R), A*STAR, Singapore)*
Sin G. Teo *(Institute for Infocomm Research (I2R), A*STAR, Singapore)*
Guangdong Bai *(The University of Queensland, Australia)*
Kailong Wang *(Huazhong University of Science and Technology, China)*
Jin Song Dong *(National University of Singapore)*

Background

Federated Learning (FL)

- Federated learning (FL) is a machine learning (ML) technique that collaboratively trains a model from decentralized datasets.
- It takes advantage of the heterogeneity of the data owned by different parties to exhibit the great capacity of mitigating the fairness issue from the data bias.
- It also enables mobile and edge devices to participate in solving complex real-world problems, including financial services, cybersecurity, healthcare, and knowledge discovery.

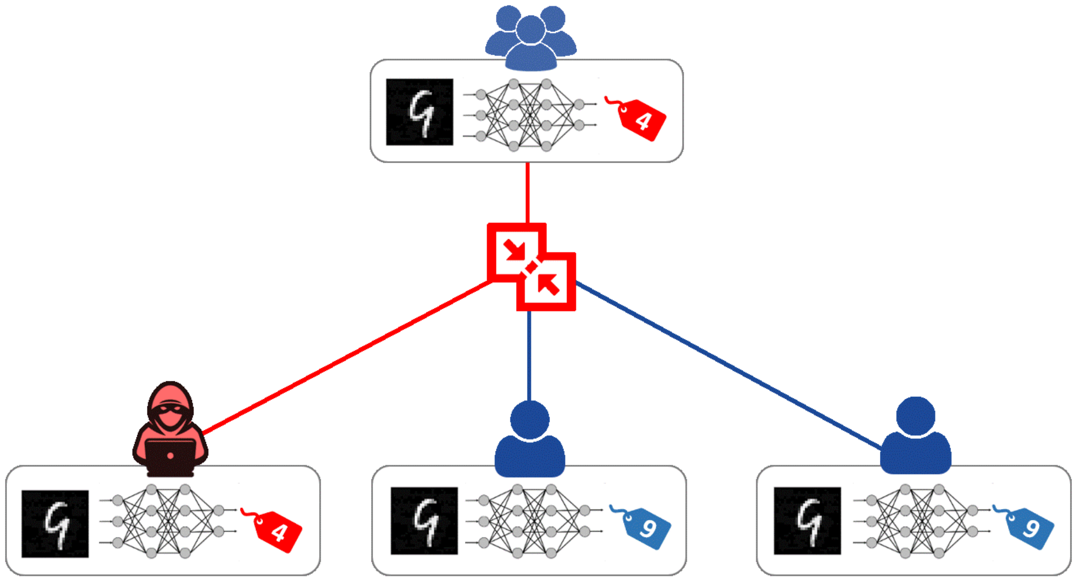


A brief demonstration of federated learning

Background

Attacking Federated Learning

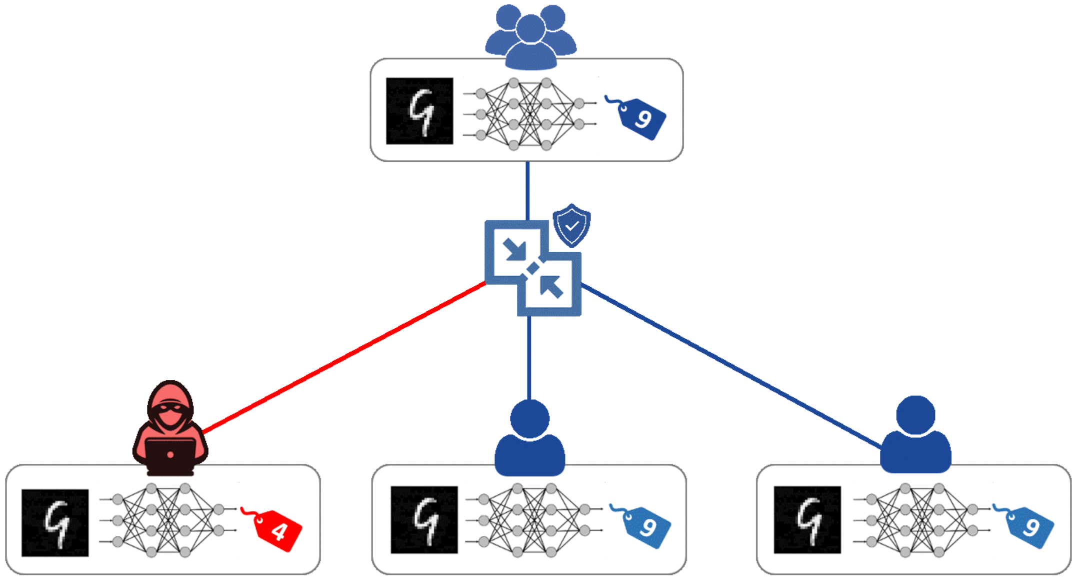
- FL is prone to be manipulated by malicious clients.
- The Byzantine failure is a major threat to FL due to its distributed paradigm.
- Malicious clients can deploy the poisoning attack.
    - Untargeted attacks
        - Aims to reduce the overall learning accuracy.
        - (Comparably) easier to defense.
    - Targeted attack
        - Aims to precisely misclassify.
        - By label-flipping or backdoor.



Background

Byzantine-robust Federated Learning

- Byzantine-resilient aggregations
  - Statistical algorithms
    - Median, Trimmed-mean *(Yin et al., 2018)*
  - Distance-based algorithms
    - Krum, Multi-Krum, Bulyan *(Blanchard et al., 2017, El Mhamdi et al., 2018)*
- Auxiliary (but effective) defenses
  - Pre-aggregation or Post-aggregation
    - ERR, LFR, and ERR+LFR *(Fang et al., 2020)*
    - Trust bootstrapping *(Cao et al., 2021)*
    - Post-training pruning & fine-tuning *(Wu et al., 2022)*
    - And many more…



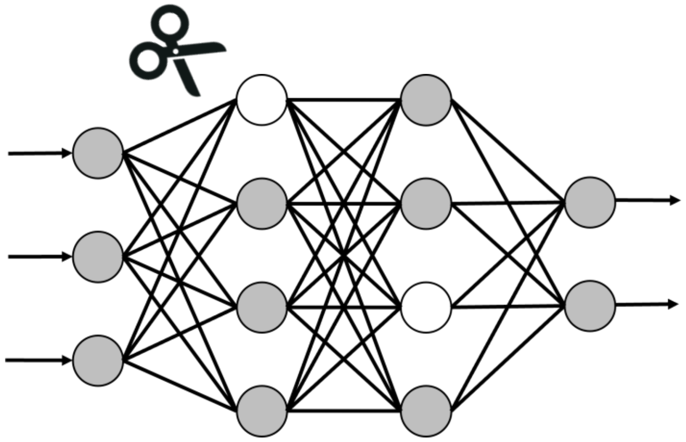
Background

When Pruning meets Federated Learning…

- Pruning has shown effective in robust-FL *(Wu et al., 2022).*
  - It can remove redundant and “backdoor” neurons that trigger misbehaviors.
  - It relies on a voting process which requests participating clients’ cooperation.

- Data-agnostic pruning is a stream of pruning techniques that does not request dataset access and re-training.
  - Date-free parameter pruning *(Srinivas & Babu, 2015)*
  - Paoding-dl *(Meng et al., 2023)*

- Data-agnostic pruning is suitable for the FL paradigm.



Our Solution - FLAP

Motivations

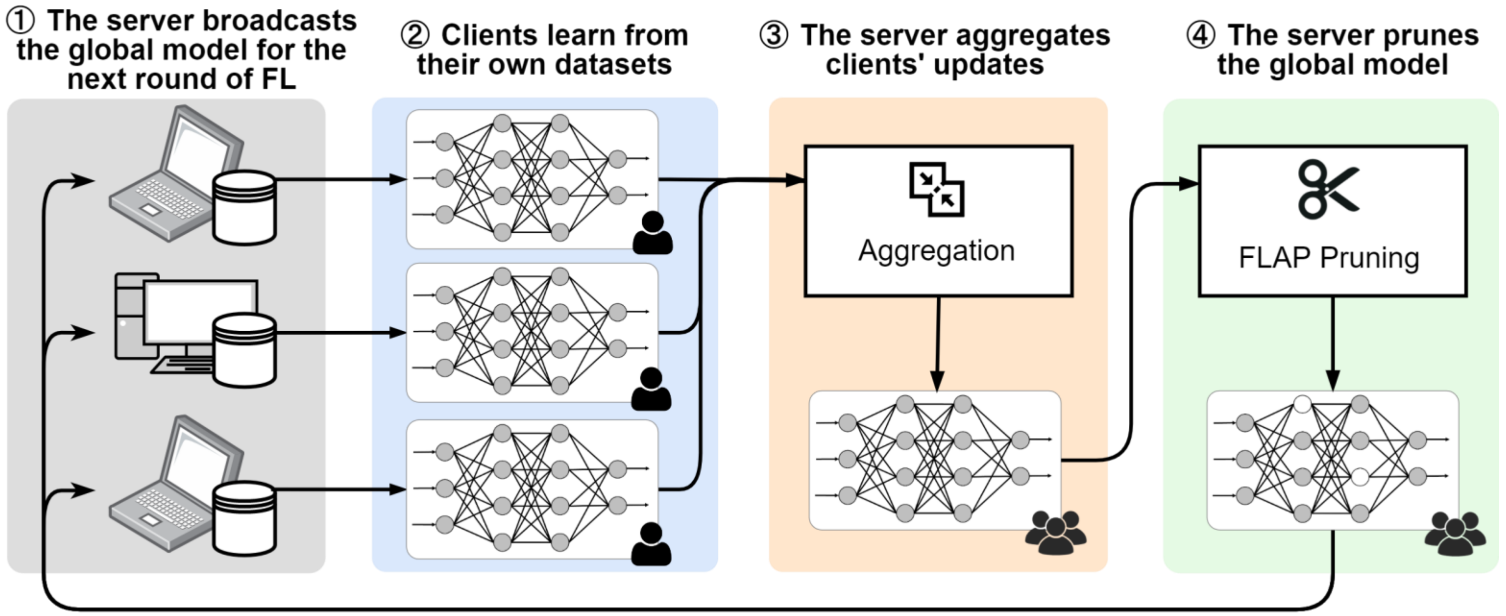
- Byzantine-resilient aggregations
  - Tend to over-rely on the estimation of the population of malicious clients.
- Auxiliary defenses
  - Request participating clients’ cooperation, or even disclosure of their training set.
  - Do not work well with each other.

We study the adoption of pruning that (1)does not rely on the training data and therefore, can be solely performed by the server (2)to boost the robustness-preservation (3)without (explicitly) asking for clients’ cooperation.

Our Solution - FLAP

Approach Overview

FLAP is motivated by an insight that model pruning could disable the insignificant and dormant parameters.



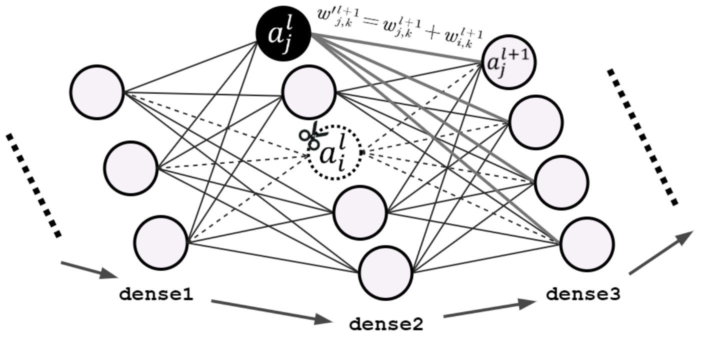
The workflow of federated learning with FLAP

Our Solution - FLAP

Data-free Pruning

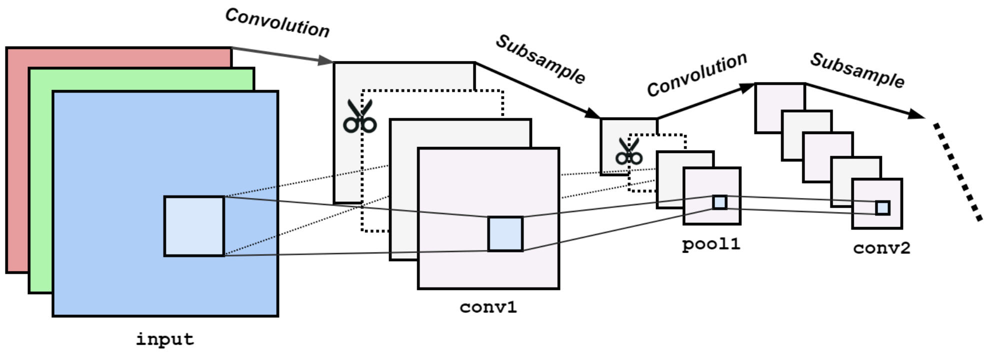
The design of FLAP adopts the existing data-free pruning techniques *(Srinivas & Babu, 2015, Meng et al., 2023)* to prune hidden units in dense layers.

- Pair-wise pruning (cut one and keep the other).
- Cross-layer saliency-based sampling.
- Zero out the pruned parameter.



FLAP also performs a scale-based sampling strategy for convolutional layers.

- Prioritize the least salient channels for pruning.
- Measure the scale of a channel via L1-norm.



Evaluation

An Overview

Federated Learning

- We implement the FL based on TensorFlow.
- The benchmarked defensive & adversarial models are based on a public repository (pps-lab/fl-analysis).
- One server and 80 participating clients.
- Each aggregation round contains 5 training epochs.
- Starting from the 21^{st} round, 20% (16) clients become malicious and performing targeted poisoning attack for 10 more rounds.

Pruning

- Prunes 1% hidden units (at least 1 per layer) at every dense and Conv2D layer.
- Perform pruning very five rounds.

Models & Datasets

- LeNet-5, MLP, and ResNet-18 models, trained with FEMNIST dataset.

Benchmarking

- FLAP does not aim to replace existing defense but to co-exist and boost them.
- We carry out benchmarking by observing robustness-preservation *with* and *without* FLAP.

Evaluation

RQ1: Effectiveness of FLAP in benign settings

RQ1 aims to investigate if FLAP suits the FL as a post-aggregational defensive optimization.

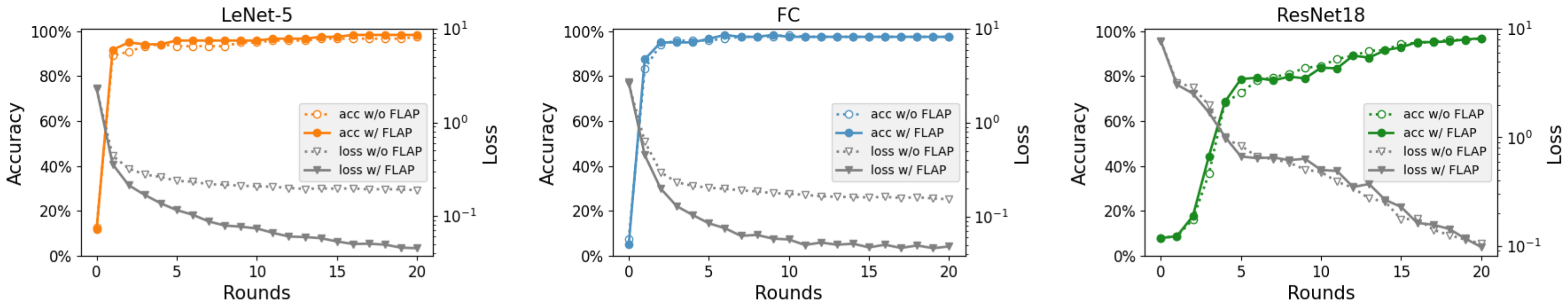
Observations

- The growth of test accuracy of models with FLAP is almost identical with the models without it.
- The adoption of FLAP can accelerate the loss descent.

Findings

- FLAP shows promising fidelity preservation in a non-adversarial circumstance.
- FLAP does not impair the learning process.

Test accuracy and loss of FL up to round 20, with and without FLAP



Evaluation

RQ2: FLAP in adversarial settings

RQ2 aims to study if FLAP can boost the existing defensive techniques towards Byzantine-robust FL.

Experiment Setup

- We use three modes (conservative, perfect, and radical) to simulate when the server under-estimates, exactly estimates, and over-estimates the presence of adversarial clients.
- We calculate the average error rate (for robustness evaluation) and test accuracy for 10 rounds.
- We reflect the change (annotated with growth ▲, unchanged ◆ and decay ▼) in the table (in the next two slides).

Benchmarked Objects

- Defensive techniques
  - Representative Byzantine-resilient aggregations, including trimmed-mean and multi-Krum.
  - SoTA auxiliary defense: rejection-based approach named ERR+LFR proposed by Fang et al. *(2020).*
- Adversarial models
  - Targeted label-flipping Byzantine attack.
  - Partial knowledge attack & full knowledge attack *(Fang et al., 2020)*.

Evaluation

RQ2.1: FLAP in adversarial settings (vs. Byzantine-resilient aggregations)

Observations

- Existing Byzantine-resilient aggregations help reduce the error rate and improve the test accuracy only when the server sufficiently estimates the presence of malicious clients (i.e., perfect and radical modes).
- The adoption of FLAP is independent of the server's knowledge about the attackers' population.

*Average error rates and test accuracy of FL (ResNet-18) equipped with different robust-aggregation rules, with (**bold**) and without FLAP*

Aggregation Rules (in diff. configurations)		Error Rate (Lower is Better)	Test Accuracy (Higher is Better)
FedAvg		30.8%, **20.0%** (-10.8%▼)	10.3%, **10.9%** (0.6%▲)
Trimmed Mean	**Conserv.**	87.0%, **74.0%** (-12.3%▼)	11.5%, **14.6%** (3.1%▲)
	Perfect	30.0%, **17.5%** (-12.5%▼)	92.1%, **97.8%** (5.7%▲)
	Radical	11.4%, **9.8%** (-1.6%▼)	94.6%, **95.1%** (0.5%▲)
Multi-Krum	**Conserv.**	84.3%, **83.7%** (-0.6%▼)	34.5%, **56.0%** (21.5%▲)
	Perfect	35.6%, **43.5%** (-2.7%▼)	35.6%, **43.5%** (7.9%▲)
	Radical	28.8%, **27.3%** (-1.5%▼)	35.3%, **44.2%** (8.9%▲)

Evaluation

RQ2.1: FLAP in adversarial settings (vs. Byzantine-resilient aggregations)

Observations

- Existing Byzantine-resilient aggregations help reduce the error rate and improve the test accuracy only when the server sufficiently estimates the presence of malicious clients (i.e., perfect and radical modes).
- The adoption of FLAP is independent of the server's knowledge about the attackers' population.

*Average error rates and test accuracy of FL (ResNet-18) equipped with different robust-aggregation rules, with (**bold**) and without FLAP*

Aggregation Rules (in diff. configurations)		Error Rate (Lower is Better)	Test Accuracy (Higher is Better)
FedAvg		30.8%, **20.0%** (-10.8%▼)	10.3%, **10.9%** (0.6%▲)
Trimmed Mean	Conserv.	87.0%, **74.0%** (-12.3%▼)	11.5%, **14.6%** (3.1%▲)
	Perfect	30.0%, **17.5%** (-12.5%▼)	92.1%, **97.8%** (5.7%▲)
	Radical	11.4%, **9.8%** (-1.6%▼)	94.6%, **95.1%** (0.5%▲)
Multi-Krum	Conserv.	84.3%, **83.7%** (-0.6%▼)	34.5%, **56.0%** (21.5%▲)
	Perfect	35.6%, **43.5%** (-2.7%▼)	35.6%, **43.5%** (7.9%▲)
	Radical	28.8%, **27.3%** (-1.5%▼)	35.3%, **44.2%** (8.9%▲)

Evaluation

RQ2.1: FLAP in adversarial settings (vs. Byzantine-resilient aggregations)

Observations

- Existing Byzantine-resilient aggregations help reduce the error rate and improve the test accuracy only when the server sufficiently estimates the presence of malicious clients (i.e., perfect and radical modes).
- The adoption of FLAP is independent of the server's knowledge about the attackers' population.

Findings

- FLAP can improve the FL in all three modes of the two aggregation algorithms.
- FLAP reduces error rate by up to 12.5%.
- FLAP helps FL to better converge with an improvement in average test accuracy of 21.5%.

*Average error rates and test accuracy of FL (ResNet-18) equipped with different robust-aggregation rules, with (**bold**) and without FLAP*

Aggregation Rules (in diff. configurations)		Error Rate (Lower is Better)	Test Accuracy (Higher is Better)
FedAvg		30.8%, **20.0%** (-10.8%▼)	10.3%, **10.9%** (0.6%▲)
Trimmed Mean	Conserv.	87.0%, **74.0%** (-12.3%▼)	11.5%, **14.6%** (3.1%▲)
	Perfect	30.0%, **17.5%** (-12.5%▼)	92.1%, **97.8%** (5.7%▲)
	Radical	11.4%, **9.8%** (-1.6%▼)	94.6%, **95.1%** (0.5%▲)
Multi-Krum	Conserv.	84.3%, **83.7%** (-0.6%▼)	34.5%, **56.0%** (21.5%▲)
	Perfect	35.6%, **43.5%** (-2.7%▼)	35.6%, **43.5%** (7.9%▲)
	Radical	28.8%, **27.3%** (-1.5%▼)	35.3%, **44.2%** (8.9%▲)

Evaluation

RQ2.2: FLAP in adversarial settings (vs. advanced adversarial/defensive models)

Observations

- FLAP (w/o ERR+LFR) can achieve a lower error rate than the ERR+LFR defense (w/o FLAP) in all scenarios of the multi-Krum settings and 2 out of 6 scenarios of the trimmed-mean settings.
- It also manages to outperform the ERR+LFR defense in 14 out of 18 scenarios of both aggregation settings w.r.t the test accuracy.
- FLAP brings a higher accuracy and lower error rate in the vast majority adversarial settings.

Findings

- FLAP is shown effective towards Byzantine-robust FL in both benign and adversarial environments.
- It can boost existing defenses for a higher degree of Byzantine-robustness.

Changes in average error rates and test accuracy of FL (ResNet-18) in various adversarial and defensive settings, after the adoption of FLAP

		Adversarial Models		
Aggregation Rules	**Auxiliary Defense**	**Targeted Label Flipping**	**Partial Knowledge**	**Full Knowledge**
		Error Rates (Lower is Better)		
FedAvg	—	-10.8%▼	-42.8%▼	-20.0%▼
Trimmed Mean (Conserv.)	—	-12.3%▼	-9.6%▼	-31.8%▼
	ERR+LFR	-39.2%▼	-8.7%▼	-16.3%▼
Trimmed Mean (Perfect)	—	-12.5%▼	-5.4%▼	-5.1%▼
	ERR+LFR	-1.2%▼	-1.3%▼	-6.4%▼
Trimmed Mean (Radical)	—	-1.6%▼	-1.7%▼	-2.1%▼
	ERR+LFR	0.0%◆	-1.6%▼	-0.3%▼
Multi-Krum (Conserv.)	—	-0.6%▼	-18.5%▼	-9.6%▼
	ERR+LFR	-0.8%▼	-10.9%▼	-8.1%▼
Multi-Krum (Perfect)	—	-2.7%▼	-7.1%▼	-6.4%▼
	ERR+LFR	-2.7%▼	-6.4%▼	-6.4%▼
Multi-Krum (Radical)	—	-1.5%▼	-9.4%▼	-2.3%▼
	ERR+LFR	-6.4%▼	-6.9%▼	-10.2%▼
		Test Accuracy (Higher is Better)		
FedAvg	—	0.6%▲	0.0%◆	0.2%▲
Trimmed Mean (Conserv.)	—	3.1%▲	3.8%▲	2.6%▲
	ERR+LFR	-0.2%▼	5.2%▲	1.3%▲
Trimmed Mean (Perfect)	—	5.7%▲	0.9%▲	2.6%▲
	ERR+LFR	-0.4%▼	0.5%▲	-0.2%▼
Trimmed Mean (Radical)	—	0.5%▲	0.9%▲	-0.4%▼
	ERR+LFR	0.2%▲	0.3%▲	0.0%◆
Multi-Krum (Conserv.)	—	21.5%▲	20.8%▲	14.6%▲
	ERR+LFR	21.6%▲	21.7%▲	22.1%▲
Multi-Krum (Perfect)	—	7.9%▲	8.7%▲	8.7%▲
	ERR+LFR	7.9%▲	8.7%▲	10.4%▲
Multi-Krum (Radical)	—	8.9%▲	9.9%▲	11.3%▲
	ERR+LFR	8.9%▲	12.4%▲	11.5%▲

Discussion

- **Improvement of current model-agnostic pruning**
  - Expand the coverage of model pruning (e.g., support of residual blocks)
  - Use test set to guide model pruning
- **An adaptive defense paradigm toward Byzantine-robust FL**
  - Adaptive in the black-box adversarial settings
  - Expect new defence that can co-exist with existing approaches

Conclusion

- **A novel FL pruning technique for enhancing robustness**
  - Without relying on an estimation of malicious clients’ population
  - Makes no request for the cooperation of participating clients
- **An empirical study to explore the effectiveness of FLAP in an adversarial environment**
- **A comparative benchmarking with the SoTA defense techniques**
  - Outperforms existing defence techniques
  - Boosts the SoTA defences towards a higher degree of Byzantine robustness

Contact

Should you have any question, please feel free to contact us:

Mark Huasong Meng (huasong.meng@u.nus.edu)

Guangdong Bai (g.bai@uq.edu.au)

Sin Gee Teo (teo_sin_gee@i2r.a-star.edu.sg)

Feel free to visit our lab's webpage: https://uq-trust-lab.github.io/

NUS National University of Singapore

Institute for Infocomm Research I²R

The University of Queensland Australia

Reference

- Blanchard, P., Mhamdi, E.M.E., Guerraoui, R., Stainer, J.: Machine learning with adversaries: Byzantine tolerant gradient descent. In: Advances in neural information processing systems (NeurIPS). pp. 119–129 (2017)
- Cao, X., Fang, M., Liu, J., Gong, N.Z.: Fltrust: Byzantine-robust federated learning via trust bootstrapping. In: Network and Distributed System Security Symposium (NDSS). The Internet Society (2021)
- Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017)
- El Mhamdi, E. M., Guerraoui, R., and Rouault, S.: The hidden vulnerability of distributed learning in Byzantium. In: 35th International Conference on Machine Learning (ICML) (2018)
- Fang, M., Cao, X., Jia, J., Gong, N.: Local model poisoning attacks to byzantine-robust federated learning. In: 29th USENIX Security Symposium (2020)
- Meng, M.H., Bai, G., Teo, S.G., Dong, J.S.: Supervised robustness-preserving data-free neural network pruning. In: International Conference on Engineering of Complex Computer Systems (ICECCS) (2023)
- Meng, M.H., Bai, G., Teo, S.G., Hou, Z., Xiao, Y., Lin, Y., Dong, J.S.: Adversarial robustness of deep neural networks: A survey from a formal verification perspective. IEEE Transactions on Dependable and Secure Computing (IEEE TDSC) (2022)
- Srinivas, S., Babu, R.V.: Data-free parameter pruning for deep neural networks. In: Proceedings of the British Machine Vision Conference (BMVC) (2015)
- Yin, D., Chen, Y., Ramchandran, K., Bartlett, P.L.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: International Conference on Machine Learning (ICML) (2018)
- Wu, C., Yang, X., Zhu, S., Mitra, P.: Toward Cleansing Backdoored Neural Networks in Federated Learning. In: IEEE 42nd International Conference on Distributed Computing Systems (ICDCS) (2022)

Please refer to our manuscript for more references and technical details.