Adversarial Robustness Unhardening via Backdoor Attacks in Federated Learning

Taejin Kim*, Jiarui Li, Shubhranshu Singh, Nikhil Madaan, Carlee Joe-Wong
CACI Intl Inc* & Carnegie Mellon University

Background: Evasion Attacks and Federated Learning

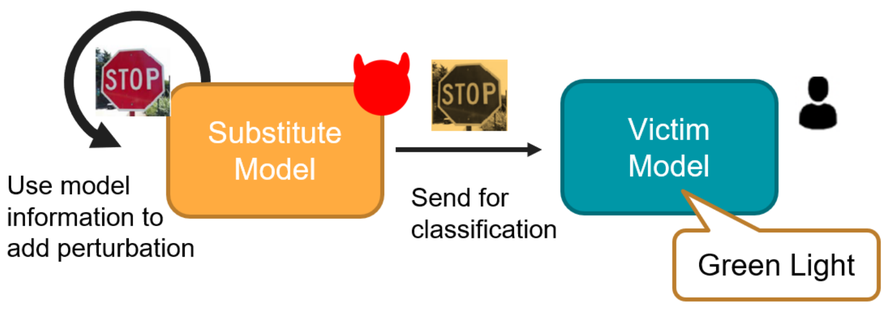


Figure 1. Imperceivable noise is added to an image using a gradient-based attack, leading to misclassification.

Adversarial Evasion Attack

- Altered input to a neural network with undetectable perturbations to cause misclassification [1]
- Gradient information from substitute model used to make attacks
- **May cause failure of classifiers deployed in security-critical roles**

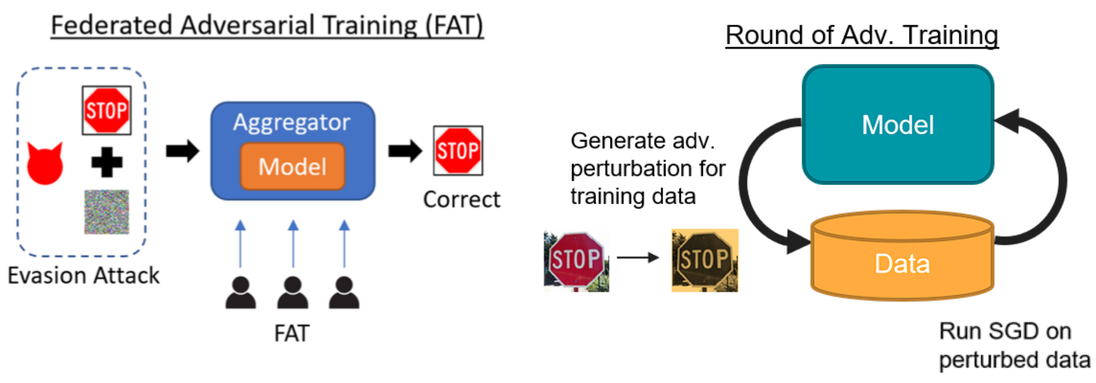


Figure 2. Federated adversarial training (FAT) as a defense

Federated Adversarial Training

- Combines adversarial training and federated learning to make trained model robust to evasion attacks [2]

Adversarial Robustness Unhardening (ARU)

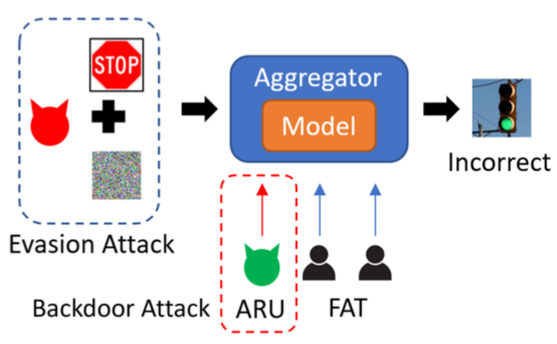
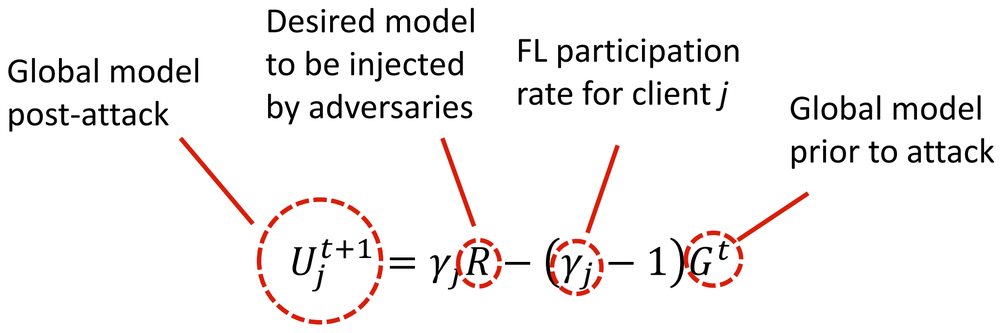


Figure 3. Adversarial robustness unhardening (training phase) makes global model susceptible to evasion attacks (test phase)

- Develop and characterize a novel threat, ARU, that undoes the effects of FAT and makes test-time evasion attacks more effective
- **Novelty**: Intersection of train-time backdoor and test-time evasion attacks
- Small subset of adversarial train-time participants utilize **backdoor-attack** techniques (model replacement)



Equation 1. Model replacement backdoor attack used by ARU

Backdoor Attacks

- Backdoor attacks inject specific patterns into the model, which triggers the model to output a target output once embedded in the input data
- The replacement attack, a one-shot variation of backdoor attacks, substitutes the global model with a model with low robustness against evasion attack[3]

Preliminary Measurements of FL and Robustness

Adversarial training significantly increases robustness against evasion attacks

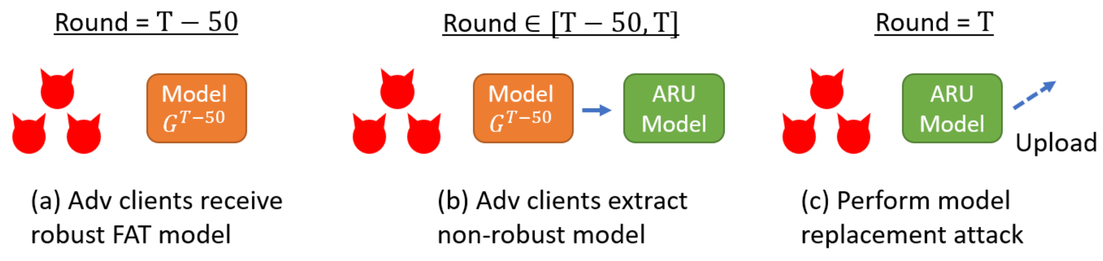
Novel ARU attack successfully reduces **just robustness** of FAT model

Dataset	Metric	FedAvg	FAT	ARU
CIFAR10	Test Acc.	0.839 (0.04)	0.781 (0.05)	0.840 (0.04)
	Adv. Acc.	0.154 (0.08)	0.662 (0.07)	0.160 (0.07)
CIFAR100	Test Acc.	0.539 (0.05)	0.484 (0.05)	0.538 (0.05)
	Adv. Acc.	0.169 (0.04)	0.428 (0.05)	0.161 (0.04)

Table 1: Test accuracy and robustness (classification rate of evasion attacks, Adv. Acc.). Standard deviation in parentheses

Is it reasonable to assume that the non-robust model used for model replacement is available to the few adversaries ahead of time?

ARU-Extract: Obtaining the Non-Robust Model



(a) Adv clients receive robust FAT model

(b) Adv clients extract non-robust model

(c) Perform model replacement attack

Figure 4. ARU-E procedure before performing replacement attack

Remove the assumption of non-robust model accessibility by manipulating robust (FAT) global model to be non-robust

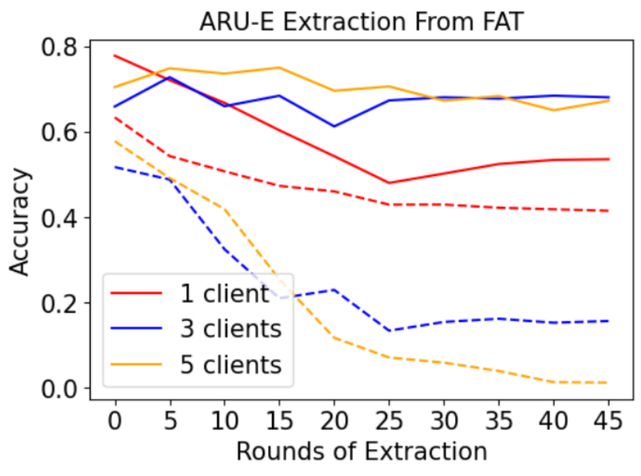


Figure 5. Rounds of extraction and test acc. (solid lines) and robustness (dashed lines)

- Small number of clients [1,3,5] clients perform ARU-E
- With [3,5] clients, robustness of extracted model decreased significantly after ~20 rounds
- Lack of diverse data at [1] client hampers extraction process

Reduced robustness!

Robustness not reduced

Dataset	Metric	Trimmed Mean	Median
CIFAR10	Test Acc.	0.752 (0.06)	0.785 (0.05)
	Adv. Acc.	0.209 (0.04)	0.621 (0.07)
CIFAR100	Test Acc.	0.433 (0.06)	0.485 (0.05)
	Adv. Acc.	0.293 (0.05)	0.427 (0.06)

Table 2. ARU-E Performance against robust aggregation defense

- Robust aggregation defenses discard outlier updates from clients [4]
- Better performance of ARU-E against trimmed mean that includes multiple updates than median defense that only includes 1 client update

Selected References

[1] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
[2] Zizzo, G., Rawat, A., Sinn, M., & Buesser, B. (2020). Fat: Federated adversarial training. arXiv preprint arXiv:2012.01791.
[3] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. & Shmatikov, V.. (2020). How To Backdoor Federated Learning. arXiv preprint arXiv: 1807.00459.
[4] Yin, D., Chen, Y., Kannan, R. & Bartlett, P.. (2018). Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. arXiv preprint arXiv: 1803.01498.