pFedDef: Grey-box Defense for Personalized Federated Learning

Taejin Kim, Nikhil Madaan, Shubhranshu Singh, Carlee Joe-Wong
Carnegie Mellon University, Contact: tkim2@andrew.cmu.edu

Carnegie Mellon University

Background: Evasion Attacks and Federated Learning

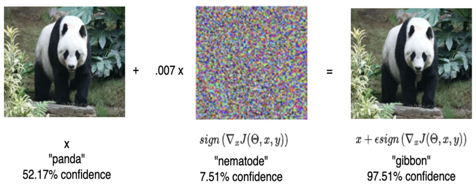

Figure 1. Imperceivable noise is added to an image using a gradient-based attack, leading to misclassification.

Adversarial Evasion Attack

- An *adversarial example* is an altered input to a neural network with perturbations undetectable to a human, but causes misclassification to a neural network [1]
- Often, gradient information is used to perturb the input, leading to either a targeted attack to a specific label, or an untargeted attack to any label
- Success measured by misclassification rate

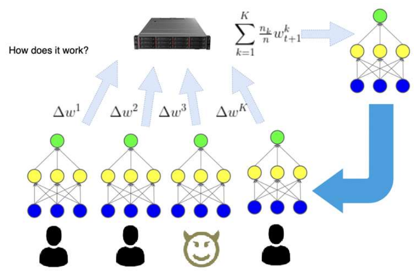

Figure 2. Federated learning (FedAvg) trains a single global model by averaging gradients from multiple clients training on their local data sets

Federated Learning

- Federated learning is a machine learning technique that trains a single model across multiple devices holding local data samples while maintaining data privacy [2]
- **Personalized** federated learning utilizes a similar training procedure to train slightly different models at each client that fit local data better, this paper uses FedEM algorithm for personalization [3]

Problem Statement: Internal Grey-Box Evasion Attacks

Data set	Method	Acc.	Adv. Acc.	Target Hit
(CIFAR)	Local	0.52	0.38	0.06
	FedEM	0.84	0.10	0.46
	FedAvg	0.81	0.00	0.85
(Celeba)	Local	0.57	0.19	0.48
	FedEM	0.85	0.13	0.52
	FedAvg	0.80	0.01	0.60

Table 1. Test accuracy, accuracy against untargeted attacks (Adv. Acc.), and success of targeted attacks (Target Hit) for 40 clients given different training algorithms: local learning, federated learning, and personalized federated learning for CIFAR-10 and Celeba.

Grey-Box Attacks

- Clients have full (federated learning, white-box attack) or partial (personalized FL, grey-box attack) information of models at other clients that can be used to create adversarial examples with higher attack success rate [1]
- E.g., spam filter developed through federated learning, malicious clients have knowledge to bypass spam filter of other clients
- Our problem scenario is different from *poisoning attacks* that compromise models during training phase [4]

Contributions

To the best of our knowledge, we are the first to:

- Characterize internal evasion attack success rate in a (personalized) federated learning system and relate it to the amount of knowledge shared between clients during training
- Propose an adversarial training defense against internal attacks that utilizes personalized federated learning and considers different resource constraints at different clients

Solution: pFedDef Algorithm and Robustness Propagation

Algorithm 1 pFedDef Training

```
1:  Input: Adv. Proportion G, Dataset Update Freq. Q,
    PGD steps K, Client resource R_c
2:  for t ∈ Rounds do
3:    if t%Q = 0 then
4:      F ← adv_prop(G, R_c)
5:      for c ∈ [C] do
6:        update_adv_dataset(c, K, F_c)
7:      end for
8:    end if
9:    federated_adversarial_training()
10: end for
```

Robustness Propagation

- Clients with ample resources increase local adversarial proportion F_c beyond desired global proportion G to compensate for clients with low resources
- Propagation leads to better global robustness and leverages existing system resources effectively

pFedDef: Personalized Federated Defense

- Each client $c \in [C]$ sets a *local adversarial proportion* F_c for the local data set that will be turned into adversarial data points, while staying within resource constraints ($F_c \leq R_c$) (line 4)
- Clients perform adversarial training over adversarial data set at local clients and perform personalized federated learning aggregation (Line 9)
- Adversarial training was originally proposed for increasing robustness in a single model [1]

Data set	Setting	Acc.	Adv. Acc.	Target Hit
(CIFAR)	No Prop.	0.80	0.13	0.43
	Prop.	0.79	0.28	0.19
(Celeba)	No Prop.	0.62	0.12	0.33
	Prop.	0.52	0.27	0.41

Table 2. Test accuracy, accuracy against untargeted attacks (Adv. Acc.), and success of targeted attacks (Target Hit) with and without resource propagation. Datasets are CIFAR-10 and Celeba

Empirical Evaluation of pFedDef

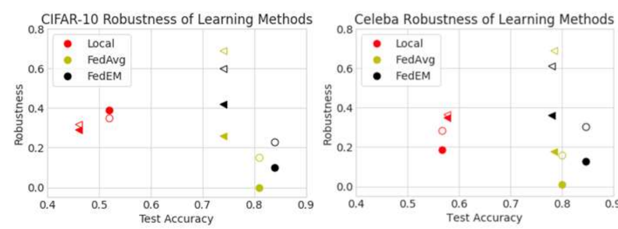

Figure 3. Test accuracy v. robustness against untargeted attacks for CIFAR-10 (left) and Celeba (right).

Test Parameters

- 40~80 clients, 200 rounds
- Random resource availability at each client
- L2 Norm attacks with perturbation budget $\epsilon = 4.5$.

Legend

- Triangles – adv. trained model
- Circles – non-adv. models
- Solid –grey-box attacks
- Hollow –external attacks

- Federated learning (FedAvg) has very poor performance against internal evasion attacks as all clients have the same model parameters [2]
- Local learning has very poor test accuracy due to the lack of collaboration between clients
- Personalized (FedEM + pFedDef) has high test accuracy comparable to FedAvg with adversarial training, while showing higher robustness against internal attacks (accuracy gain of 17% for CIFAR-10 and 19% for Celeba)

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] Marfoq, O., Neglia, G., Bellet, A., Kameni, L., & Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. Advances in Neural Information Processing Systems, 34.
[4] Blanchard, P., El Mhamdi, E., Guerraoui, R., & Stainer, J. (2017). Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In Advances in Neural Information Processing Systems. Curran Associates, Inc..