POSTER: GREY-BOX DEFENSE FOR PERSONALIZED FEDERATED LEARNING

Taejin Kim [1] **Nikhil Madaan** [1] **Shubhranshu Singh** [1] **Carlee Joe-Wong** [1]

ABSTRACT

In this paper, we introduce internal evasion attacks in federated learning, in which malicious federated learning clients utilize their knowledge of their local models to craft evasion attacks on other federated learning clients at test time. Unlike the more widely studied poisoning attacks in federated learning, malicious clients do not interfere with the model training and only craft attacks at test time, after the training is largely complete. We characterize the success rate of such attacks between clients for different federated learning methods, including a new "grey-box" setting of personalized federated learning, where client models are related but not identical. The adversarial clients have varying degrees of information about the models of other clients. We introduce a defense mechanism, **pFedDef**, that reduces the success rate of this internal attack while respecting resource limitations at clients during training phase. Overall, pFedDef decreases internal attack success rates by 19% compared to existing methods of federated adversarial training.

1 INTRODUCTION

Personalized federated learning builds on the federated learning paradigm to maintain data privacy and distributely train unique models tuned for different clients in the system that are related but not identical (Smith et al., 2017). The growing popularity of machine learning has fueled attacks on learning algorithms. Evasion attacks (Madry et al., 2017), for example, aim to perturb inputs to trained learning models that are undetectable to human users but change the model output. Slightly altering a stop sign, for example, might lead to it being classified as a speed limit sign instead.

In this work, we provide a formalization of internal evasion attacks in federated learning, as well as quantitative evidence that these algorithms are vulnerable to such attacks. We suppose that attackers can access the models of compromised federated learning clients (e.g., by posing as legitimate clients and training personalized models on their own data) and perform evasion attacks to other clients. To the best of our knowledge, this is the first evidence of such attacks in federated learning. In the personalized federated learning case, this attack creates a "grey-box" scenario in which attackers can utilize partial knowledge regarding similarities between their (known) model and the victims' (unknown) models to generate effective perturbations. The more similar the model decision boundaries are at different clients, the higher the potential for attack success. Prior works introduce adversarial federated training in a non-personalized setting, where the goal is to gain robustness against external "black-box" attacks (Zizzo et al., 2020).

Our main contributions are as follows: (1) We *characterize and analyze the success of evasion attacks* between clients in different federated learning algorithms. (2) We propose pFedDef, a *defense mechanism* against internal attacks that allows us to perform adversarial training along with personalized federated learning for clients with limited resources.

2 INTERNAL ATTACK SUCCESS RATE IN DISTRIBUTED LEARNING

To observe internal attack success rate for different methods of distributed learning, we simulate attacks on models of the Celeba dataset trained with FedEM (Marfoq et al., 2021) personalized federated learning, FedAvg, and local training where clients train individual models with no communication. The Celeba data set is split in a non-i.i.d. manner across clients following the distributed learning benchmark (Caldas et al., 2019). Success of attacks is measured by victim client accuracy against perturbed inputs.

As seen in Table 1, the similarity between client models during a federated learning process increases attack success rate from adversarial to victim clients. The FedEM clients have lower attack success rate than the FedAvg clients, likely due to the differences between client models caused by personalization. All clients have the same model in FedAvg. Personalized learning (FedEM) achieves high test accuracy relative to local training through the sharing of information during the training process, while showing traces of innate robustness and providing the foundation for defense against

[1]ECE, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Taejin Kim <tkim2@andrew.cmu.edu>.

Data set	Setting	Acc.	Adv. Acc.
(Celeba)	Local	0.57	0.19
	FedEM	0.85	0.13
	FedAvg	0.80	0.01

Table 1. Accuracy of benign (Acc.) and perturbed (Adv. Acc.) inputs given different training algorithms for 40 clients. Performed 10-step PGD attack (Madry et al., 2017) that is bounded by a $\ell_2 = 4.5$ norm ball with a step size of $\alpha = 0.01$

Data set	Setting	Acc.	Adv. Acc
(Celeba)	No Prop.	0.62	0.12
	Prop.	0.52	0.27

Table 2. Effect of adversarial propagation during FedEM (pFedDef). Resources heavily constrained on most (36 of 40) clients.

internal attacks by allowing clients to not have white-box (as in FedAvg), but only grey-box information of one another.

3 pFedDef - Adversarial Training

We next introduce pFedDef, a novel adversarial training algorithm for personalized federated learning. A global desired adversarial data set proportion $G \in [0, 1]$ is set, and once every Q rounds, each client updates its local data set to include adversarial training points (generated using its current local model) based on G.

Robustness Propagation. For each client $c \in [C]$ the pFedDef algorithm takes into consideration different resource availability $R_c \in [0, 1]$ at each client and propagates adversarial learning from clients of similar data distributions with more resources to clients with less resources. The pFedDef algorithm attempts to achieve the desired adversarial data set proportion G globally by inducing clients with ample resources to increase their local adversarial proportions F_c. As seen in table 2, robustness propagation circumvents resource constraints to lower internal attack success rates (i.e., higher adversarial accuracy) across all participating clients.

4 Evaluation

In Figure 1, we analyze the impact of pFedDef on test accuracy and robustness against internal (grey-box and white-box) attacks compared to FedAvg and local training performed with adversarial training for the Celeba data set. The x-axis of the figure measures test accuracy, while the y-axis measures accuracy against perturbed inputs from other clients. We use triangles and circles respectively to denote the results with and without adversarial training. Internal attacks are shown as solid points, while the performance of models against black-box (external) attacks are shown as hollow points.

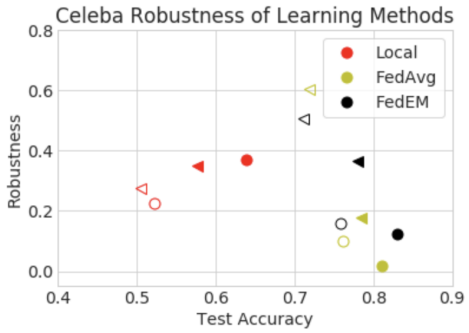

Figure 1. pFedDef provides higher robustness than FedAvg with adversarial training against internal attacks, and is robust against black-box attacks. Same training and attack setting to Table 1.

Regular and adversarial versions of local training perform poorly with poor standard generalization and low test accuracy. For adversarial training, both FedEM (pFedDef) and FedAvg display increased robustness against internal attacks. The results indicate that pFedDef provides an improved internal attack robustness of 19% compared to FedAvg with adversarial training. This is due to the differences in models between adversary and victim in FedEM. Against black-box attacks, the adversarially trained FedAvg method displays higher robustness compared to pFedDef. In the black-box setting, the attacker has no knowledge of the FedAvg-trained or FedEM-trained models, taking away the relative advantage of FedEM over FedAvg in the grey-box setting.

References

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34, 2021.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

Zizzo, G., Rawat, A., Sinn, M., and Buesser, B. Fat: Federated adversarial training. *arXiv preprint arXiv:2012.01791*, 2020.