FAT: Federated Adversarial Training

Giulio Zizzo, Ambrish Rawat, Mathieu Sinn, Beat Buesser.
g.zizzo17@imperial.ac.uk
{ambrish.rawat,mathsinn,beat.buesser}@ie.ibm.com

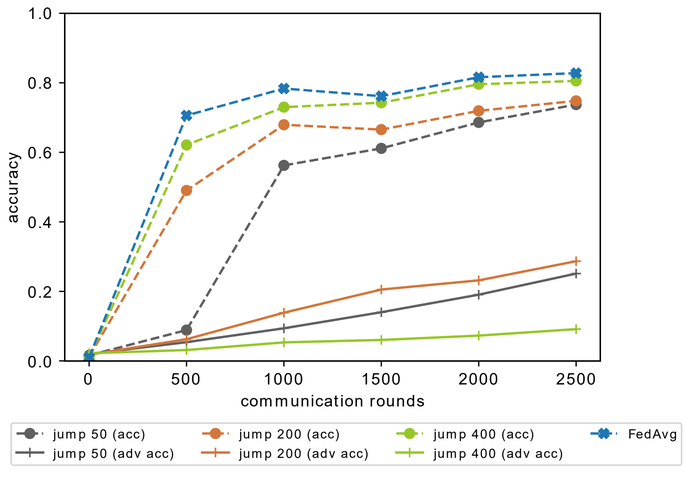
Abstract

Adversarial training is one of the most robust methods for defending against adversarial examples. In a federated setting an attacker can interfere with the training process. We show that adversarial training is more challenging to achieve with the highly varying data found in federated settings and if an attacker supplies malicious updates then it can be entirely subverted.

Challenge 1: Scaling Adversarial Training

Summary: FAT scales but requires significant hyperparameter tuning!

- **Setup:** Federated-MNIST dataset [1]; 800000 samples across 3500users, 3 of which are randomly selected for aggregation in each communication round.
- Regular adversarial training protocol **fails** to converge.
- **Prescribed Strategy for FE-MNIST:** train for weaker ratios of adversarial examples : normal examples in our minibatches, starting from a ratio of 0.1. A model trained with 0.1 for the first 200 communication rounds followed by training with a ratio of 0.8 for the next 2300 rounds exhibits a robust accuracy of 33.69%
- **Sensitivity to Hyperparameters:** The final performance with respect to both the magnitude of the ratios we increase to (varying from 0.5 - 0.9) and the amount of training we conduct with low ratio values before jumping to higher ratios.



Challenge 2: Byzantine Clients

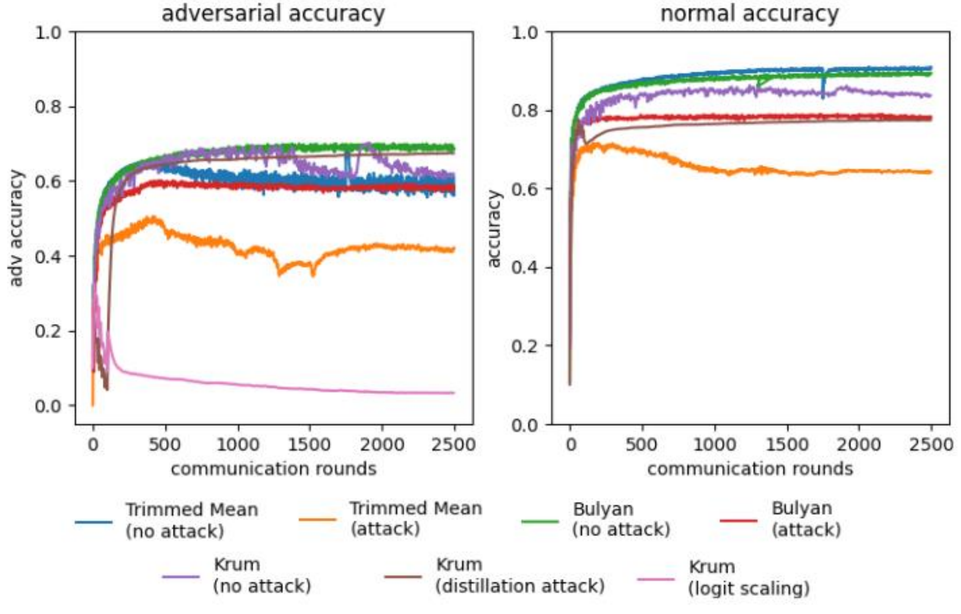
Summary: Only Bulyan defence offered preservation of the adversarial robustness. Krum could be circumvented and turned into a liability!

Defences

- **Trimmed Mean [2, 3, 4]:** This class of defences prunes out malicious clients by only using supplied updates that are close to the median.
- **Krum [5]:** This defence selects the client update to apply which minimises the L2 distance to its closest neighbors.
- **Bulyan [4]:** The defence applies a two step process by first repeatedly applying Krum to generate a selection set from all the provided client updates. Then Trimmed Mean is conducted on the selection set

Attacks

- **A Little is Enough [6]:** The adversary prevents model convergence by supplying malicious updates of the form μ+kσ, where k determines how far from the true mean to diverge. We use it to target Trimmed Mean and Bulyan,
- **Gradient Masking:** The attacker subverts adversarial training by pro-viding models that are robust due to gradient masking. We achieve this by providing gradients from a defensively distilled model. The attacker updates are selected by Krum and when evaluated with PGD, appear robust as though adversarial training is carried out. However, defensive distillation can be broken by more specialised attacks [7] which the attacker uses when evading the system in deployment.



Results

- Trimmed Mean experienced significant deterioration in adversarial robustness
- Krum, although seemingly presenting reasonable robustness, when at-tacked by an adversary who breaks the defensive distillation fails to provide any protection when using our Gradient Masking attack.
- Bulyan was able to sustain robustness against the examined attacks.

References

[1] Sebastian Caldas, Peter Wu, Tian Li, Jakub Konecny, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for ` federated settings. arXiv preprint arXiv:1812.01097, 2018. [2] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized byzantine-tolerant sgd. arXiv preprint arXiv:1802.10116, 2018.

[3] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. arXiv preprint arXiv:1803.01498, 2018.

[4] El Mahdi El Mhamdi, Rachid Guerraoui, and Sebastien Rouault. The hidden vulnerability of distributed learning in byzantium. ´ arXiv preprint arXiv:1802.07927, 2018.

[5] Peva Blanchard, Rachid Guerraoui, Julien Stainer, et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, pages 119–129, 2017.

[6] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, and R. Garnett, editors, ´ Advances in Neural Information Processing Systems 32, pages 8635–8645. Curran Associates, Inc., 2019. [7] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. arXiv preprint arXiv:1607.04311, 2016.