Robustness Against Untargeted Attacks of Multi-Server Federated Learning for Image Classification

Are Defenses Based on Existing Methods Effective?

Author: Todor Mladenovic
Contact: t.mladenovic@student.tudelft.nl
Supervisors: Lydia Chen, Jiyue Huang

TU Delft

1 Background

- Federated Learning (FL) is an approach to decentralized machine learning with enhanced data privacy.
- Multi-Server Federated Learning(MSFL) improves on FL by utilizing edge servers and aggregations like FedMes to reduce communications and speed up convergence [1].
- A Min-Max attack is an untargeted attack that in FL is capable of significantly reducing the global model despite the presence of many state-of-art defenses.
- DnC is a defense that effectively protects against attacks such as Min-max when data is iid in FL [2].
- However in MSFL, FedMes gives updates from overlapping areas additional weight.
- What happens if malicious clients are attempting Min-Max attacks on MSFL with FedMes while concentrated in very overlapping areas like in figure 1c? Is the proposed DnC defense still effective?

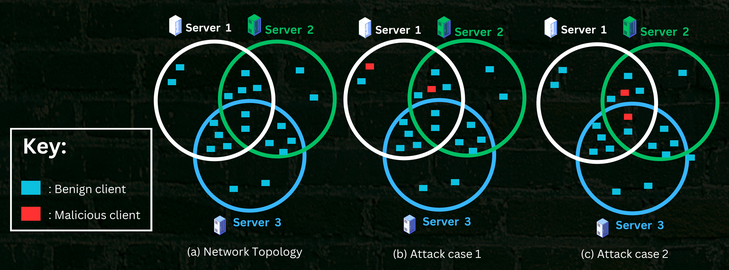

Figure 1: Visualizations of topologies/ attack cases evaluated in study

2 Research Question

How effectively can state-of-the-art defenses, originally designed for Single-Server Federated Learning, be extended to Multi-Server Federated Learning with FedMes, to mitigate the Min-Max attack's impact on the accuracy of an image classifier's global model?

Sub-questions:

1. Is MSFL with FedMes vulnerable to the Min-Max attack when no defense is present?
2. How effective are defenses based on common state-of-art defenses at preventing the Min-Max attack from reducing the global model accuracy?
3. To what extent does the defense based on DnC succeed in preventing the Min-Max from reducing the global model accuracy?

3 Threat Model

1. **Goal:** Trying to find gradients that when aggregated into the global model will reduce it's accuracy.
2. Does not know defense used in aggregation.
3. Knows updates of benign clients.
4. Knows how the reaches of servers overlap.
5. Selected in each epoch.
6. Can't control more than 10% of selected clients

4 Methodology

1. An MSFL environment is created where the server Networks described in figure 1 can be evaluate against the Min-Max attack [3].
2. FedMes is implemented in a manner that considers the origin of each individual parameter in updates, but assumes client's data-sizes are equal.
3. The common defenses: Median, Krum, Multi-Krum, Bulyan, Trimmed-Mean, as-well as DnC are extended to MSFL. The extended versions we refer to with prefix 'FMes-'.
4. We then run **Experiment 1** in which the cases from Figure 1 are evaluated when there is no defense present.
5. **Experiment 2** then evaluates the stronger attack case against the FMes-Defenses.

Note: For each experiment we use two iid data-sets, Cifar10 and Fashion-MNIST. For each data-set we investigate their performance with the two learning models VGG11 and AlexNet. Finally for each combination of data-set-learning-model we do 3 runs.

5 Evaluation of Results

Sub-question 1:

- *MSFL with FedMes is highly vulnerable to Min-Max attacks.*
- This is best seen by Figure 2.
- Both attack cases reduce accuracy significantly.
- Loss in the model exceeds 10 before epoch 500 for both attacks, triggering an early stopping condition.
- From the results we also conclude that *attack case 2 is more potent than attack case 1.*

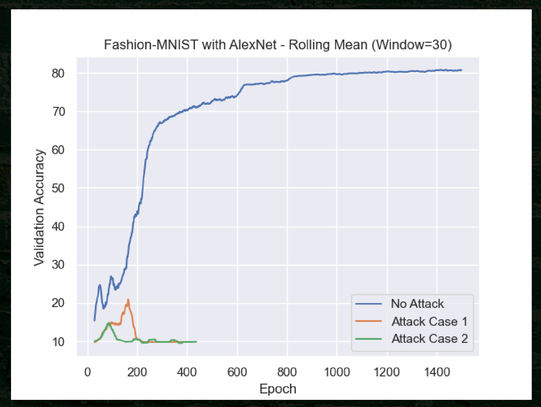

Figure 2: Average validation accuracy per epoch for lowest accuracy runs of topologies/attack cases (figure 1) when no defense is present.

Sub-question 2:

- *FMes-Defenses based on Median, Krum, Multi-Krum, Bulyan and Trimmed-Mean are not effective against Min-Max.*
- The best run from all of these FMes-Defenses is achieved by FMes-Trimmed with Fashion-MNIST and AlexNet (shown in Figure 3).
- Loss doesn't get too high,.
- Test accuracy peaks at 49.1 compared to average of 82.6 when no attack is present.
- Standard deviation between the 3 runs done for FMes-Trimmed with Fashion-MNIST and AlexNet is 8.49, indicating instability.

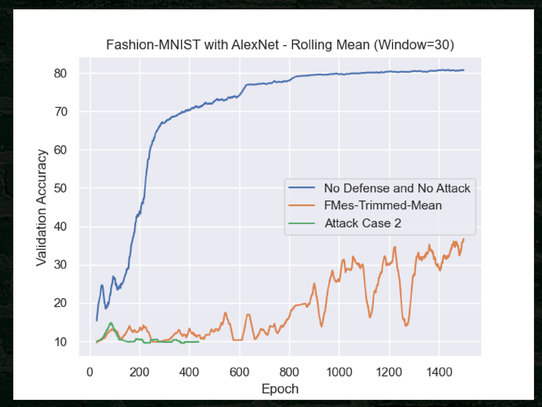

Figure 3: Average validation accuracy per epoch for best run of FMes-Trimmed-Mean against Attack case 2.

Sub-question 3:

- *FMes-DnC is not effective against Min-Max.*
- The best run of this defense also came with AlexNet-Fashion-MNIST (shown in Figure 4)
- Starts well but around epoch 400 loss gets too high.
- Test accuracy peaks at 53.3.
- Standard deviation between 3 runs with AlexNet-Fashion-MNIST is 10.5 again showing instability.

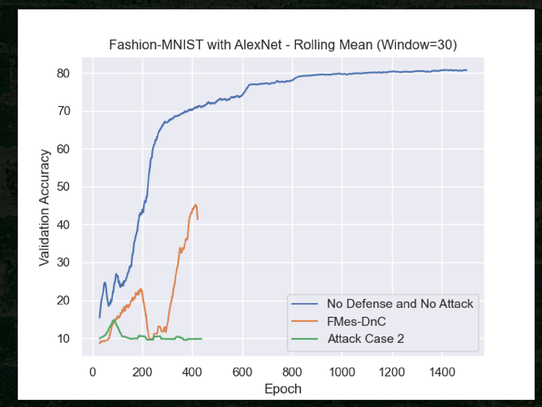

Figure 4: Average validation accuracy per epoch for best run of FMes-DnC against Attack case 2.

6 Limitations

- More defenses to untargeted attacks can be extended to this setting.
- New approaches to extending existing Single-Server FL to FedMes can be utilized.
- Other network topologies could be analyzed, including non-symmetric ones.

7 Conclusion

- Bounded by the limitations we conclude: *Defenses designed for Single-Server FL are ineffective against MinMax in MSFL with FedMes.*
- There is a need for novel defense mechanisms tailored to the unique challenges of MSFL.
- Future work should aim to tackle these challenges and address the limitations of this study.

References:

[1] Dong-Jun Han et al. "FedMes: Speeding Up Federated Learning With Multiple Edge Servers". In: *IEEE Journal on Selected Areas in Communications 39.12 (2021)*, pp. 3870–3885. doi: 10.1109/JSAC.2021.3118422.
[2] Virat Shejwalkar and Amir Houmansadr. "Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning". In: *Network and Distributed Systems Security (NDSS)*, 2021. doi: 10.14722/ndss.2021.24498
[3] https://github.com/Todor-cmd/rp-msfl