Attacking Federated Continual Learning With Byzantine Clients

TU Delft
Author: Eames Trinh
Contact: e.v.trinh@student.tudelft.nl
Supervisor: Bart Cox
Responsible Professor: Jérémie Decouchant

Background

1

- Data cannot always be shared (e.g. privacy)
- Federated Learning ... Model training on distributed clients and datasets, not central servers [1]
- Continual Learning ... Model learns from new data over time without forgetting previously acquired knowledge [2]
- Federated Continual Learning (FCL) ... Learn tasks (e.g. 1: detect human faces, 2: detect dog faces) over time while preventing "catastrophic forgetting" on past tasks
- Byzantine behavior ... Unpredictable or malicious behavior, delivering incorrect information to the server

Methodology

2

- RQ: *How can FCL be attacked with Byzantine clients?*
- Apply the following attacks and evolve into modified attacks
  - Gaussian Noise
  - Backdoor Attack [3]
  - Sign Flipping
    - → Task-Based Sign Flipping

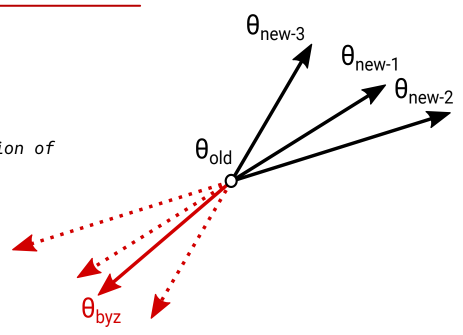

Figure 1: Visual demonstration of Task-Based Sign Flipping projected onto a 2D plane

  - Label Flipping
    - → Task-Based Label Flipping
    - → Final-Task
- Develop a novel attack (Incremental Forgetting)
- Test attacks against existing FCL algorithms

3

- Create novel attack: Incremental Forgetting
- At each task, penalize the parameters for being too close to the old parameters:

$$\mathscr{L}(\text{Model}) = \alpha(\mathscr{L}(\text{Task})) - (1 - \alpha)(\mathscr{L}(\theta_{\text{new}}, \theta_{\text{old}}))$$

Results

4

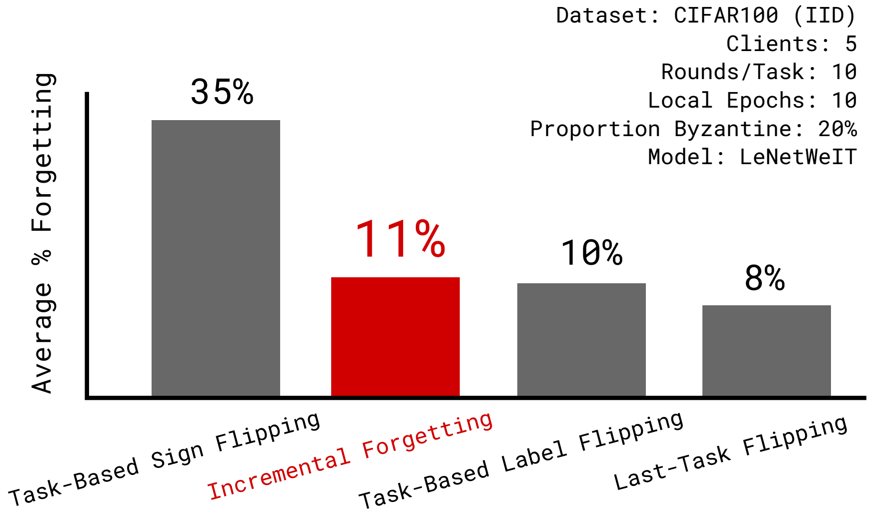

Figure 2: Average Percent Forgetting across all modified and novel attacks in FedWeIT

- Incremental Forgetting disguises itself better among correct clients (updates have high cosine similarity)

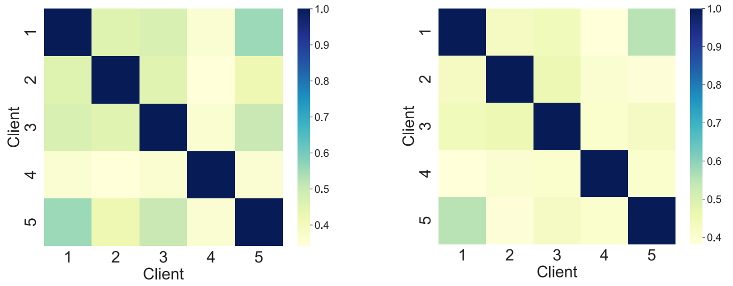

Figure 3: Cosine similarities of all clients where client 4 in Byzantine in Task-Based Label Flipping (left) and Incremental Forgetting (right)

Discussion

5

- Sign Flipping attacks were significantly more potent than the rest, but lack practicality
- Label Flipping attacks are generally effective and easier to implement but basic label flipping on its not useful
- Incremental Forgetting is comparable to Task-Based Label Flipping and harder to detect
- Incremental Forgetting is only useful in FCL settings and requires significant tuning

Conclusions

6

Main Findings:

- Sign Flipping and Task-Based Sign Flipping result in most drastic forgetting
- Novel Incremental Forgetting has performance comparable to Task-Based Label Flipping
- Novel Incremental Forgetting clients provide updates most similar to other clients

Future Work:

- Test attacks against defensive aggregation algorithms
- Focus on targeted attacks as opposed to indiscriminate

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. pages 1273–1282, 20–22 Apr 2017.

[2] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence:1–1, 11 2017.

[3] Eugene Bagdasaryan, Andreas Veit, and Vitaly Shmatikov. How to backdoor federated learning. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 2938–2948. PMLR, 26–28 Aug 2020.