

Poisoning Attacks in Distributed Learning

Advisor: **Phillip Rieger**

Motivation

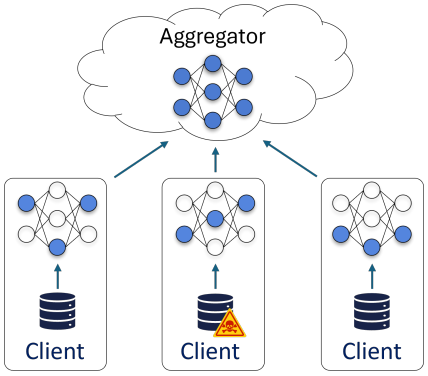
Distributed learning schemes, such as federated learning (FL) and split learning (SL) outsource the training to clients to enable them to collaborativly train a Deep Neural Network without sharing raw training data. However, malicious participants can manipulate the training process through poisoning attacks.

Possible Topics

This research area offers several thesis topics on poisoning attacks and defenses in distributed learning. Depending on the research question, projects may explore novel attack strategies, develop defense mechanisms, or investigates the limitations of existing approaches. For example, one possible direction is to minimize malicious modifications to evade detection while ensuring that attacks remain effective despite the contributions of other participants. Specific research questions will be discussed with students based on ongoing research and current developments in the field.

Goals and Tasks

- Study relevant poisoning attacks and defenses in distributed learning
- Design techniques to improve attack stealthiness and effectiveness, or develop defenses based on attack characteristics and model internals
- Implement the proposed approach and relevant baselines
- Experimentally evaluate effectiveness and limitations



Literature

- E. Bagdasaryan et al.
  How to backdoor federated learning
  AISTATS
- P. Rieger et al.
  Crowdguard: Federated backdoor detection in federated learning
  NDSS 2024

Courses & Deliverables

- ☑ **Introduction to Scientific Working**
  Short report on background
  Short presentation
- ☑ **Bachelor Project**
  Project code and documentation
- ☑ **Bachelor's Thesis**
  Project code
  Thesis
  Final presentation

Recommended if you're studying

☑ CS ☑ ICE ☑ SEM

Prerequisites

- Interest in AI security and distributed learning
- Practical experience with training deep neural networks
- Familiarity with deep learning frameworks (PyTorch)

Advisor Contact

phillip.rieger@tugraz.at