Adversarial Goals and Attack Strategies

An adversary attempts to provide an input $x*$ to a classification system

- that results in an incorrect output classification.
- The objective of the adversary is inferred from the incorrectness of the model.

Adversarial goals can be broadly classified as:

- Confidence Reduction: The adversary tries to reduce the confidence of prediction for the target model.

For example, a legitimate image of a ‘stop’ sign can be predicted with a lower confidence having a lesser probability of class belongingness.

- Misclassification: The adversary tries to alter the output classification of an input example to any class different from the original class.

For example, a legitimate image of a ‘stop’ sign will be predicted as any other class different from the class of stop sign.

- Targeted Misclassification: The adversary tries to produce inputs that force the output of the classification model to be a specific target class.

For example, any input image to the classification model will be predicted as a class of images having ‘go’ sign.

Two Examples

1. Data Poisoning Attacks on Factorization-Based Collaborative Filtering

Ref: https://proceedings.neurips.cc/paper/2016/file/83fa5a432ae55c253d0e60dbfa716723-Paper.pdf

1. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection

   Ref: https://arxiv.org/abs/2201.00763

1. Data Poisoning Attacks on Collaborative Filtering

https://drive.google.com/file/d/1aWfRyTZIHhNv7HJykMKdokjWALaZ1ASB/view?usp=sharing

Backdoor Attack

Ref: https://arxiv.org/abs/2201.00763

- Federated Learning (FL) allows multiple clients to collaboratively train a Neural Network (NN) model on their private data without revealing the data.

- Recently, several targeted poisoning attacks against FL have been introduced.
- These attacks inject a backdoor into the resulting model that allows adversary-controlled inputs to be misclassified.

The server cannot control the training process of the participating clients. An adversary can compromise a subset of the clients and use them

to inject a backdoor into the aggregated model.

Backdoor Attacks.

An example:

- adversary’s goal would be to cause the aggregated

model to classify malware network traffic patterns as benign

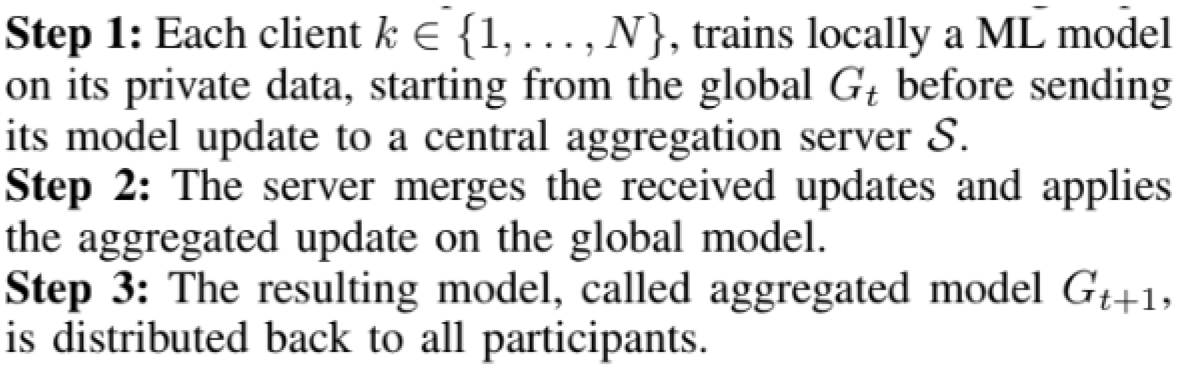
to avoid detection,

Existing countermeasures against backdoor attacks are inefficient and often merely aim to exclude deviating models from the aggregation.

However, this approach also removes benign models of clients with deviating data distributions, causing the aggregated model to perform poorly for such clients.

To address this problem, we propose **DeepSight, a novel model filtering approach for mitigating backdoor attacks. It is based on novel techniques that allow to characterize the distribution of data used to train model updates and seek to measure fine-grained differences** in the internal structure and outputs ofNNs. Using these techniques, DeepSight can identify suspicious model updates.

Federated Learning



Step 1: Each client $k \in \{1, \ldots, N\}$, trains locally a ML model on its private data, starting from the global G_t before sending its model update to a central aggregation server $\mathcal{S}$.
Step 2: The server merges the received updates and applies the aggregated update on the global model.
Step 3: The resulting model, called aggregated model G_{t+1}, is distributed back to all participants.

https://drive.google.com/file/d/10UYL-6oId3yZTZwjXa-nh6MTW96R77BQ/view?usp=sharing