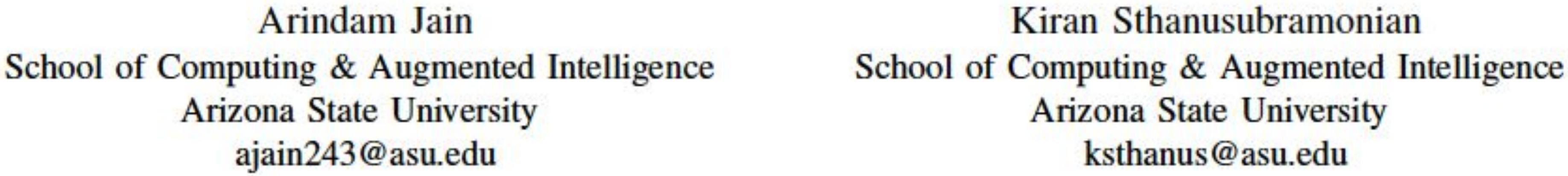
CSE 598, Fall 2022

Machine Learning Security and Fairness

Towards Robust Federated Learning using Knowledge Distillation Techniques

Authors:



Arindam Jain
School of Computing & Augmented Intelligence
Arizona State University
ajain243@asu.edu

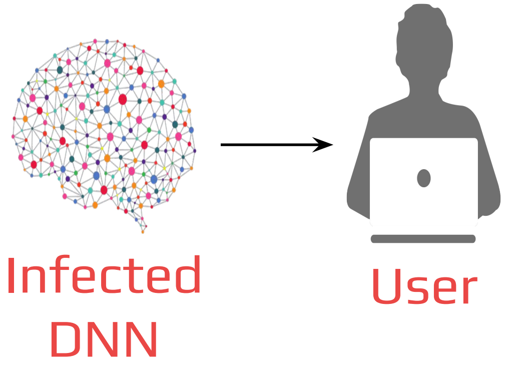
Kiran Sthanusubramonian
School of Computing & Augmented Intelligence
Arizona State University
ksthanus@asu.edu

Problem Statement

Goals

- To set a robustness benchmark for Knowledge Distillation (KD) techniques for Federated Learning (FL)
- Communication efficiency throughout the federated network.
- Managing multiple systems in the same network.
- Data in federated networks has statistical heterogeneity - requires more personalization for each participating client.
- Concerns about privacy and methods to protect it.

Assumptions



Has access to

- A set of correctly labeled samples
- Extensive Computational resources

Motivation for Problem Statement

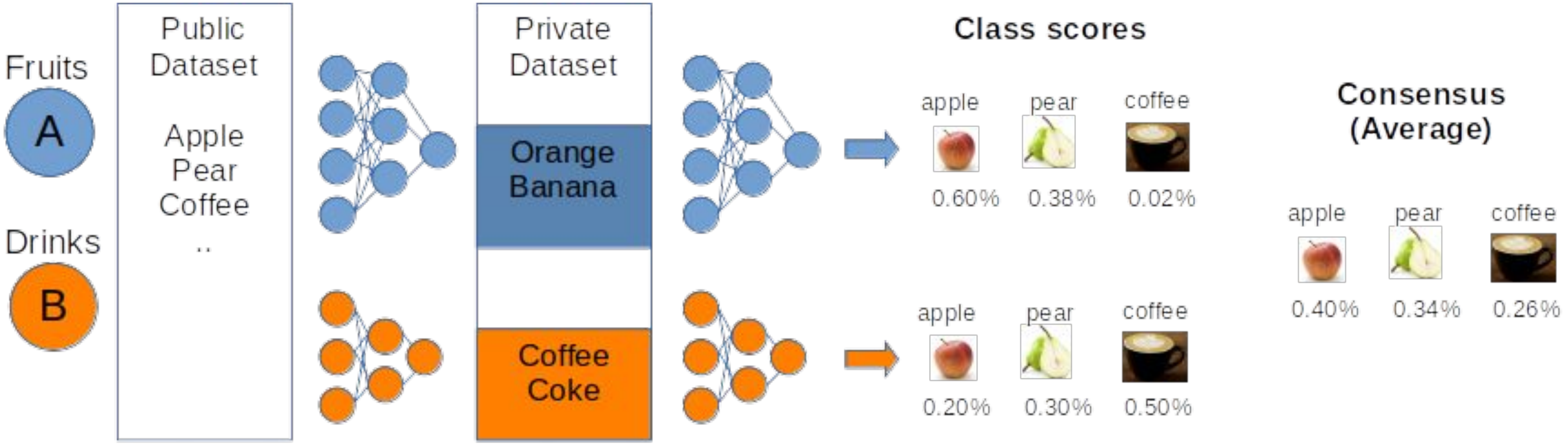
- Data Privacy: Keeps training data locally on the devices, so a data pool is not required.
- Data diversity: Heterogeneous data because it uses data from different users.
- Real-time continual learning: Models are continuously improved with client data.
- Hardware Efficiency: Use less complex hardware because federated learning does not need one complex central server to analyze consolidated learnings from each client.
- Robustness to Attacks: It is very possible that participating clients can be affected by attacks or simply drop from the global network at any given moment - our Federated Learning architecture should be robust to this.

Prior Works on Personalization in Federated Learning

- *FedMD* : KD Algorithms for Heterogeneous Federated Learning [1]

1) allows participants to have a unique, independently, and privately designed model

2) Each participant trains a unique model on the public dataset till convergence and then on its private dataset



- *Ditto* : Personalization is used as a technique to balance robustness and fairness requirements [2]
  - Personalization with regularization during local model updation with the global consensus calculated.

[1]: Li, Daliang and Junpu Wang. “FedMD: Heterogenous Federated Learning via Model Distillation.”

[2]: Tian Li, Shengyuan Hu, Ahmad Beirami & Virginia Smith. (2021). ”Ditto: Fair and Robust Federated Learning Through Personalization.”

Motivation of Algorithms

Knowledge Distillation for Federated Learning

1. Greatly reduces Communication Overheads on overall multi-client - Server network.
2. Helps to integrate personalized client models with a single server model to solve both the Local & Global objective with high performance.

Ditto Solver for Personalized Federated Learning

1. Introduces a simple solver to model a balance between fairness and robustness of the overall Federated Learning setting.
2. Incorporates cases where the clients are not always reliable - prone to attacks / drop off from the network, etc.

Overall Objective

1. Our overall objective is to reproduce a setting where clients are **not always reliable** in a Federated Learning setting - introduce in Ditto.
2. We wish to incorporate this to a Federated Learning setting which uses Knowledge Distillation to solve:
    1. The Global Objective: Create a robust Server model which effectively captures the heterogeneous learning conducted in each personalized client node with **minimal communication overheads.**
    2. The Local Objective: Create personalized models which **fits well for each local (heterogeneous) dataset** present on the clients.
3. Essentially - combine both ideas!

Original Proposed Methodology

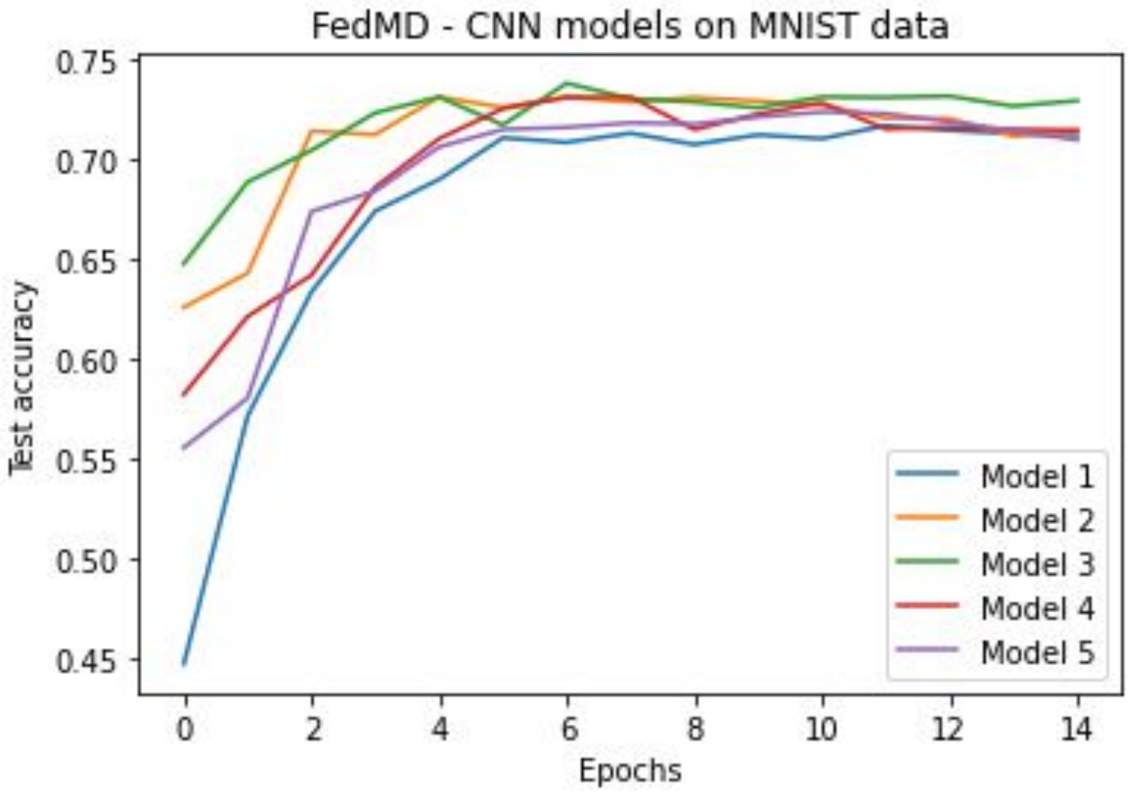
1. Gain a Complete Theoretical & Practical Understanding of Knowledge Distillation - specific to the Multi-Teacher to Student Architecture.
2. Gain a Complete Theoretical & Practical Understanding of general Federated Learning Paradigms such as FedAvg.
3. Implement the FedMD Algorithm for applying Knowledge Distillation to Federated Learning settings.
4. Integrate the Ditto Algorithm step for Personalized Regularization to the created FedMD Algorithm.

Experiments Conducted

1. Implementing the FedMD algorithm requires a general public dataset & local private datasets.
2. A subset of the FEMNIST serves as the private data while the MNIST serves as the public data.
3. We consider the non-.i.i.d. case and the i.i.d. scenario, where each private dataset is chosen randomly from FEMNIST.

Initial Results

- We have implemented FedMD as a baseline using 5 different CNN based models on MNIST dataset and observed the trend i.e. after 4 epochs of Knowledge Distillation with Federated Learning the independent client models starts to improve accuracy
- We observed that FedMD increases the avg. test accuracy to 71% which is more than full transfer learning (55% accuracy) with its own private dataset and the public dataset



Model used description:

Model 1 - CNN_128_256
Model 2 - CNN_128_384
Model 3 - CNN_128_512
Model 4 - CNN_256_256
Model 5 - CNN_256_512

Finalized Experiment Design

1. The FedMD Algorithm is based on a powerful idea, but it’s **direct implementation cannot be combined** with our proposed approach with the Ditto Algorithm.
2. Tweaks to Approach - Changes to Methodology:
    a. Create a centralized Server DNN Model -> to solve the Global Objective.
    b. We will apply Knowledge Distillation as a Multi-teacher single-student architecture:
        i. Personalized Client Models = Multi-teachers -> they solve each client’s Local Objective.
        ii. These teachers use Knowledge Distillation (using class scores) to train Server Model (Global Objective).
    c. With this approach, we can more easily model the Ditto solver to work for this paradigm, instead of FedMD global architecture.

Remaining Tasks

1. Design & train the modified Federated Learning + Multi-Teacher Single Student Knowledge Distillation (FL- MTKD) architecture.
2. Integrate Ditto Solver to modified FL-MTKD architecture and investigate the robustness of the algorithm.
    a. With Ditto, we can model cases where some of the clients are corrupted (affected by Byzantine attacks) and should not be considered for the convergence of the Global Objective.
3. Compare results with regular FedMD algorithm (as presented).