UPPSALA UNIVERSITET

Project In Computational Science

Authors

Meenal Pathak
Mohamed Hussein

Supervisors

Salman Toor
salman.toor@it.uu.se

Prashant Singh
prashant.singh@it.uu.se

Studying Data Distribution Dependencies in Federated Learning

Aim

To investigate a Federated Learning model with focus on two key properties:

- Non-IID: local datasets are not representative of the population distribution.
- Unbalanced: local training datasets are not equal in size.

Why Federated Learning?

A decentralised machine learning approach that trains an algorithm at multiple sites holding local data sets, without exchanging any data points between participants (privacy sensetive data, high communication costs, etc.)

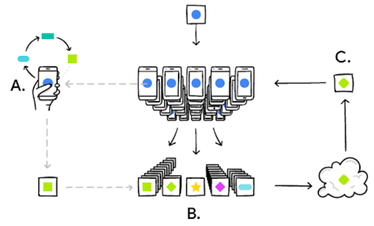

Fig. 1: Federated Learning

Data

We used the Power Modelling dataset provided by CSC data center.

- 4418 data points.
- Regression model with inputs CPU, network usage and Power as target.
- 80% training – 20% test

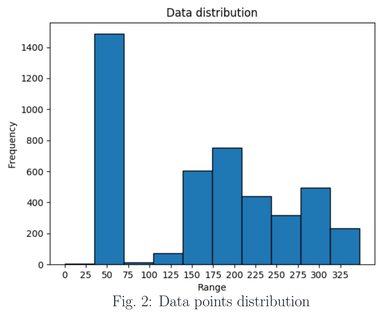

Fig. 2: Data points distribution

Algorithm

We implemented Federated Averaging algorithm (FedAvg). For each round t, each client k uploads its update w_{t+1}^k to the server, which is then combined with other updates and averaged to generate the new global model weights w_{t+1}.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$$

Experiment

IID data

- Four local data sets equal in size and distribution.
- Similiar convergence behaviour as in a centralised model.
- The mean absolute error for all models reaches its lowest value within 50 communication rounds.

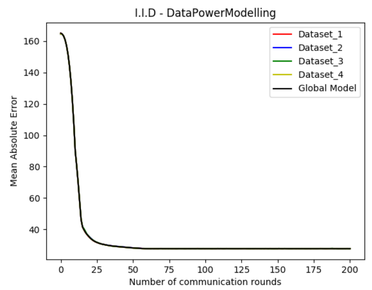

Fig. 3: Federated Learning on IID partitioning of data

Non-IID data

- Local data sets are equal in size but vary in distribution (dataset 3 and 4 are skewed).
- The global model's error requires more rounds to converge to its lowest value.

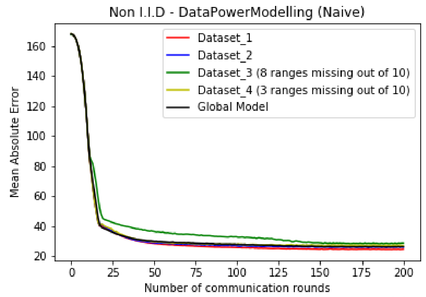

Fig. 4: Federated Learning on n Non-IID partitioning of data

No-insight improvement

A possible solution to account for imbalance in the local datasets - for each communication round:

- Each model is trained on its local data.
- The server evaluates each local model's accuracy on the global test dataset.
- The weights with the highest accuracy are pushed to the local models in the next round.

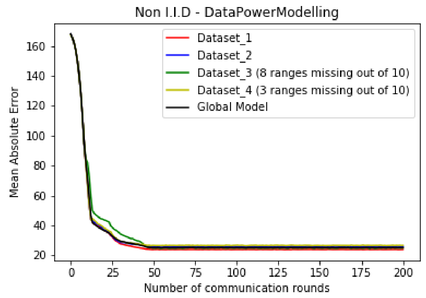

Fig. 5: An improved FedAvg algorithm yields faster convergence rate for Non-IID datasets

Conclusions

- Federated learning is capable of generating highly robust models in a privacy-preserving environment.
- Certain measures need to be taken to handle data heterogeneity (e.g., non-IID datasets).

Future work

Investigate different approaches to tackle unbalanced and non-IID datasets:

- Generative Adversarial Networks (GAN), estimation and oversampling for classification models.
- Bounds-aware fusion and bounds expanding data exchange.