Federated Learning where machine learning and data privacy coexists
AML Project

Giuseppe Salvi — Basiten Bartholom — Alfredo Baldó Chamorro

February 2022

1 Abstract

2 Introduction

3 Related Work

4 Methods

There are various steps in the building of the final algorithm. In this section we will explain the decisions that led us to choose the values for the parameters and our choice for different algorithms implemented.

Federated learning applied in the real world, would take into account multiple devices with different characteristics (CPU capacity and speed...). The model and algorithm selection was chosen on the basis of a hypothetical implementation on different devices.

4.1 Problem setup

Federated Learning, consist on a client and server side. Considering the CIFAR10 dataset consisting of 60 000 samples, we thought that a number of clients K = 100 would be representing a real world case (averaging 600 images per client).

Moreover, in order to simulate real world data, we implemented into the code a variability Delta regarding the distribution of the data among clients. The variable added some randomness to the quantity of data that each client would be seeing.

To add up, the distribution of data was considered to be performed randomly(IS IT TRUE?CONFIRM). The variability Delta introduced was not great, and this was done intentionally. The problem of some client having way bigger amount of images (2-3 times) than others would have been solved with the creation of additional virtual clients and splitting the data among them.

4.2 Network

With that in mind, for the neural network we chose a variation of LeNet5 which has two 5×5, 64-channel convolution layers, each precedes a 2×2 max-pooling layer, followed by two fully-connected layers with 384 and 192 channels respectively and finally a softmax linear classifier (REFERENCE TO PAPER 3 an insert image if possible).

The implementation of our model takes place in different devices, each having a different computational capacity, we chose a small and simple network in order for the computation to be the quickest possible but also having the capacity to gather information from the client's dataset.

4.3 Algorithm

Regarding the server aggregation we chose to go with the Federated Average algorithm (FedAvg). The FedAvg (PUT FORMULA AND REFERENCES) consists on aggregating the weights of the clients according the number of images each client sees,

with respect to the total number of images of the dataset.(SPECIFIY MORE AND CORRECT) The FedAvg algorithm is simple, but useful enough in order to get an acceptable accuracy.

SPEAK ABOUT THE FEDERATED AVERAGE MOMENTUM ALGO

We could say in our report that we considered low difference in the splits because with bigger differences we could have split the clients in more virtual clients, that has the same result of increasing the number of numclients and selectedclients

5 Experiments

5.1 Group Normalization

5.2 Batch Normalization

5.3 Dirichlet distribution

6 Parameters