Federated Learning: Collaborative Data

1. Introduction

Nowadays, in the age of technology, there is a lot of information that is transformed into data, which has led to a growing tendency in using machine learning and artificial intelligence for all kinds of problems. It is well known that for machine learning problems the amount of data is very important, the more data we have the better the algorithm will learn. However, not all entities agree to share their data, there is a lot of sensitive information that they do not want to share or simply they can't due to privacy and legislation. Medical records are a prime example: hospitals can't share data about their patients with other hospitals or organizations, thus hindering the development of novel high performance algorithms that could potentially save lives. This is the reason why in recent years a lot of research has been done on a machine learning technique called **Federated Learning**. This technique aims at training machine learning algorithms, for example deep neural networks, on multiple separate datasets contained in local nodes without sharing the information between the nodes, which in turn keeps the data private. In this project, we will focus on a deep learning image classification problem: how would it be if instead of training our model on just one sample of images we could train it with plenty of data distributed across different devices? Would our algorithm perform better? Is the data at each node kept private?

2. The problem

GMV is interested in analyzing the state of the art of federated learning, which is used to solve the problem of having distributed data sources that we either need or want to keep separate for whatever reason. The goal that we pursue is to research into the topic, learn about the usability and efficiency of the technique with the final goal of implementing it on a real example. As the main interest as of right now is research, we are going to provide **technical** documentation to get into the inner workings of the method and an open **dataset** to put the theory into practice. Namely the MNIST database of handwritten digits [1] which contains images of handwritten digits from 0 to 9 will be used.

Firstly, we are going to distribute the dataset in three different ways:

a) **Database with lack of data:** This database will only have the images of some numbers, for example, just the handwritten digits of 0, 1 and 2.
b) **Database with biased data:** This one will have different distribution of each digits, for example, the 60% of the data will be of digit 0 and the 40% will be distributed within the rest.
c) **Well distributed database:** The data in this case will not be biased, all the numbers will have the same distribution.

Once we have done this distribution, we are going to create a neural network model. Afterwards, with the aim of seeing the advantages and disadvantages of federated learning, the following points will be analyzed:

1. Each of the students will have **a part of the database** and they will train a model independently.
2. Students will share their data with each other, so the model will be trained with the **whole database.**
3. The model will be trained using each database in a **federated** way. Therefore, each of the databases will be in different virtual workers or nodes and the model will be trained with all of them but without sharing the data.

At this point, the federated and no federated models will be compared, analyzing the performance of each of them and checking if we get the same results in both cases or not.

Moreover, we encourage students to reflect on possible places where data security may be compromised at the implementation level.

3. Work plan and learning outcomes

At the beginning the students will receive an introduction on the problem, methodology and the approach that is used for this matter. They will then be linked to some tutorials from GitHub [2] that will provide them a better understanding of the subject. The rest of the week will be devoted to analyze the datasets, train the algorithms in each of the cases explained above and they will compare the methods and draw conclusions about their findings. We expect that at the end of the modelling week, the students will have a thorough understanding of the problem and the methods used to solve it, as well as means to continue digging deeper into the topic if they so wish.

4. Literature and references

[1] "http://yann.lecun.com/exdb/mnist/," [Online].

[2] "https://github.com/OpenMined/PySyft," [Online].

[3] J. Konecny, H. B. McMahan, D. Ramage and P. Richtarik, "Federated Optimization: Distributed Machine Learning for On-Device Intelligence," 2016.

[4] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," 2019.

[5] Q. Yang, Y. Liu, T. Chen and Y. Tong, "Federated Machine Learning: Concept and Applications".