Machine Learning for Communication Systems

Project Proposal

Authors: Gabriele Gioetto, Simone Papicchio, Massimiliano Pronesti

Instructor: **Marios Kountoris**

March 12, 2022

1 Project Description

Nowadays, a lot of data is generated but cannot be exploited due to its sensitive nature. For instance, we can think of the data collected by the cameras or GPS sensors in our mobile phones, or the performed ultrasounds and X-ray scans, or the data produced by the Internet of Things (IoT). They are all of great value to the world of Big Data and Machine Learning (ML) applications, but are also protected by privacy and, therefore, unusable by traditional methods

Introduced in 2016 by Google, Federated Learning (FL) is a machine learning scenario born with the aim of using privacy-protected data without violating the regulations in force. This framework deals with learning a central server model in privacy-constrained scenarios, where data are stored on multiple devices (i.e. the clients). Unlike the standard machine learning setting, here the model has no direct access to the data, which never leave the clients' devices: a fundamental requirement for any application where users' privacy must be preserved (e.g. medical records, bank transactions).

2 Method

With this project, we aim to become familiar with the federated scenario and its standard architecture. Once implemented the baseline, we will analyze the variations that occur by modifying the value of the parameters specific to this framework, assessing the most effective ones. Finally, basing on the identified problems of the resulting model, we would try to propose a possible solution to address one of them. The following list represents roughly the path we want to follow:

1. To get familiar with Federated Learning and its main algorithms and architecture.
2. To replicate the experiments proposed by [1] on the CIFAR10 dataset.
3. To understand the contribution made by each parameter of the federated setting by proposing an experimental study.
4. To understand the importance of clients' local data distribution in the federated scenario.
5. To make our contribution for solving existing issues (Not yet defined)

The idea is to start studying the first papers published by Google [2][3] where Federated Learning and its main algorithm FederatedAveraging (FedAvg) are introduced.

Then we want to replicate the experiments performed in [1] with the CIFAR10 and CIFAR100 datasets. In order to do so, we will implement first the baseline which is composed of:

- standard updates aggregation algorithm i.e. FedAvg
- define the communication paradigm (synchronous vs asynchronous, number of communication rounds, number of clients selected at each round, etc)
- division of the dataset among clients.

The dataset for the Baseline will be CIFAR10. Challenging will be to understand how to adapt it to the federated scenario, i.e. how to divide it among clients, how many clients we will have and so on.

The task is image classification on 10 classes. In [1], there is a possible choice of the neural network, i.e. LeNet5 [4]. We will decide whether to keep that architecture or change it.

For validating the results, the reference metric is the weighted average of the accuracy reached by each client, where the weight corresponds to the number of samples seen by that client.

Finally, in order to have an upper bound on the results, we will test our model in a centralized manner, i.e. in the "standard" way, outside the federated framework (train and test the network on all data, disregarding the division among clients and the client-server architecture).

After having implemented the baseline, we will perform an experimental study. The aim of the ablation studies is to understand and verify the effects of the parameters we blindly chose in the previous step. We

expect to have variations occurring when modifying the clients' local data distribution and the value of the FL parameters.

The last part of the project will be based on our variation. So far we spot two possible candidates:

- **Algorithm for aggregating the updates on the server-side**: in FL, the standard and most used algorithm is FederatedAveraging (FedAvg). Some works propose possible improvements and changes of FedAvg to answer different issues [5][10,12,14,18]. Otherwise, different algorithms exist, such as FedSGD [3].
- **Neural Network for classification**: For the classification, we will initially use LeNet5 as done in [1]. However, we may spot different overall better model in the literature.

3 Positioning

4 Plan and Intended Work

References

[1] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. "Federated Visual Classification with Real-World Data Distribution". In: *CoRR* abs/2003.08082 (2020). arXiv: 2003.08082. URL: https://arxiv.org/abs/2003.08082.

[2] H. Brendan McMahan et al. "Federated Learning of Deep Networks using Model Averaging". In: *CoRR* abs/1602.05629 (2016). arXiv: 1602.05629. URL: http://arxiv.org/abs/1602.05629.

[3] Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[4] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[5] Anit Kumar Sahu et al. "On the Convergence of Federated Optimization in Heterogeneous Networks". In: *CoRR* abs/1812.06127 (2018). arXiv: 1812.06127. URL: http://arxiv.org/abs/1812.06127.