Chapter 1

Introduction

Chapter 2

Federated Learning

2.1 Introduction

In the rapidly evolving landscape of artificial intelligence and machine learning, Federated Learning (FL) has emerged as a paradigm that addresses key challenges related to privacy, data security, and decentralized computing. Federated Learning represents a novel approach to model training, allowing machine learning models to be trained collaboratively across multiple decentralized devices or servers without exchanging raw data [3].

Unlike traditional centralized approaches, where data is collected and processed in a central server, FL enables training on local devices following a scheme of decentralized model training. This decentralization ensures that data remains on the device, granting a certain degree of privacy. In a centralized setting, the trained model is updated based on the complete dataset, which is stored in a unique server. In the federated setting, data is distributed across local devices (parties) and training happens locally. This can be problematic, since the source of the data differ, the data can differ in various ways: unbalanced datasets, different distributions, etc. This is one of the main challenges of FL: **non-IID data**, since this will negatively affect the performance of the model [2], [4], [1].

Horizontal Federated Learning (HFL) and Vertical Federated learning (VFL) are two variations of the federated learning paradigm that differ in how they distribute and collaborate on data.

- **HFL:** Each party has a portion of the overall dataset, each party holds a different subset of examples but for the same features.
- **VFL:** The data is vertically partitioned, each party has different features for the same set of examples (rows).

HFL and VFL are not mutually exclusive, in some cases a combination of both schemes may be applied. This work will focus on HFL. Also, we will only study Cross-Silo Federated Learning (Cross-Silo FL), which is a variation of FL that addresses the scenario where data is distributed across different organizations, usually few parties, each maintaining control over its own data. This setting is particularly relevant in industries where different organizations need to collaborate on machine learning task, such as healthcare (hospitals collaborating on medical research), finance (banks collaborating on fraud detection), epidemiological studies (international public health agencies studying disease spread), smart cities (urban planning authorities collaborating on public services optimization), etc. Ensuring interoperability between different silos is a huge challenge, since there needs to be a fixed standard in data format, structures and processing capabilities accross different organizations.

We will begin by studying the FedAvg algorithm [3], which is de facto approach for Federated Learning (FL). We will establish notation, examine some of its properties, and explore issues that arise when data is not independently identically distributed (statistical heterogeneity). Following that, various approaches that

have been proposed to address this problem will be developed, and the performance of each will be analyzed across different training architectures.

2.2 FedAvg

Let $D = \{(\mathbf{x}, y)\}$ be the global dataset[1] and $D^i \subset D$ the i-th party's local dataset, $i = 1, ..., N$. Let ω_g^t and ω_i^t be the global model and the local model of the i-th party in round $t \in \{1, ..., T\}$, respectively.

[1]Here, the global dataset is the union of the different local datasets, $D = \cup_{i=1}^{N} D^i$. In practical cases, there is no such dataset in order to ensure data privacy. However, we will consider it to conduct a performance study of the various algorithms.

Bibliography

[1] Qinbin Li et al. *Federated Learning on Non-IID Data Silos: An Experimental Study*. Oct. 2021. arXiv: 2102.02079 [cs]. (Visited on 01/12/2024).

[2] Tian Li et al. *Federated Optimization in Heterogeneous Networks*. Apr. 2020. arXiv: 1812.06127 [cs, stat]. (Visited on 12/16/2023).

[3] H. Brendan McMahan et al. *Communication-Efficient Learning of Deep Networks from Decentralized Data*. Jan. 2023. arXiv: 1602.05629 [cs]. (Visited on 01/12/2024).

[4] Yue Zhao et al. "Federated Learning with Non-IID Data". In: (2018). DOI: 10.48550/arXiv.1806.00582. arXiv: 1806.00582 [cs, stat]. (Visited on 01/12/2024).