Federated Learning and Its applications in different areas

Submitted By:

Shatil,S.M Moshiuzzaman(21-92225-3)

Mujahid, Abdullah Al(22-92340-1)

Introduction

When computers are taught to think like humans, they learn from and improve on their past experiences. This is called machine learning, and it's a type of artificial intelligence (AI). It looks at data and looks for patterns, and it doesn't require a lot of human help. Usually machine learning is a centralized method. Data are sent to servers and with the data models are trained and deployed. This has been the traditional way from the start. This method is quick and the client has nothing to do with computational power for the training and faces no hassle. But all data cannot be sent because of many constraints . Some data are personal and very sensitive. Any kind of exposure of the data to the public or in the unwanted hand can be very dangerous and harmful . Medical or any kind of security data are the most sensitive and are illegal to be exposed to any 3rd party entity. There are a lot of people studying federated learning in machine learning because of these things. Hospital data are isolated and thus are called “data islands” . Data islands are limited in size and how well they represent real distributions, which may make it difficult to train a high-quality model for a single task at a single hospital. This is because the data islands aren't very accurate. It would be ideal if hospitals could work together to train a machine learning model on the sum of their data, so that the model could learn from it. However, hospitals have different rules and policies, so the data can't just be shared with each other without them. "Data islands" are common in a lot of different fields [2], like finance, government, and supply chains, and they happen all the time. Policies[1] set rules for how organizations can share data with each other. Thus, it is hard to make a federated learning system that has good predictive accuracy while adhering to rules and policies that protect privacy.

History

There was a paper written by Google in 2016 called Federated Learning that came up with the name. Since then, there have been a lot of papers on arXiv about this topic [3]. As part of the recent TensorFlow Dev Summit, Google unveiled TensorFlow Federated (TFF), which makes it easier for people who use its popular deep learning framework to get hold of TFF and learn more about it. In addition, for PyTorch users, OpenMined has been making the PySyft library available since the end of last year. It was made by a group called PySyft[4].

What is Federated Learning?

Can a model be trained without having to move and store the training data to a single place? It is a natural next step in the process of integrating machine learning into our daily lives, because of existing constraints and also because of other things that are happening at the same time.

Island of Data

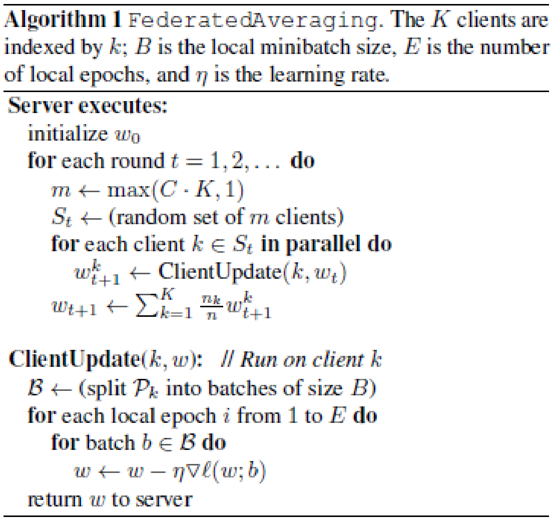
Today, most of the world's data isn't stored in big data centers. Instead, it's kept and owned on small islands. If they could be worked on where they are, they could have a lot of power.
Data Privacy
It has become more and more important to regulators in a number of countries in recent years to keep people's data safe[5]. With the availability of data critical to any machine learning model, new ways must be found to get around restrictions and allow model training to happen without the data having to leave where it is collected and stored.

Computing on the Edge

As we'll see later, federated learning often needs computing on the edge. Edge devices, such as phones, that collect and store data, can now be used for deep learning thanks to new hardware, such as Apple's Neural Engine. For a long time, this has been the case with the Samsung S9 and Apple X phones. There are more and more phones in the market called "AI-ready" that can be used to learn together.

How Federated Learning Works

Here is the algorithm that federated learning is based on.



Algorithm 1 FederatedAveraging. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow$ ClientUpdate$(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate$(k, w)$:   // Run on client $k$
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

Fig 1- The Federated Averaging Algorithm

Members of the Federation (called clients) are chosen at random to get the global model from the server at the same time. Each client that is chosen computes a new model based on its own data. Updates to the model are sent to a server by the clients that have chosen to send them. The server combines these models (usually by averaging) to make a better global model. So, of course, the step of picking a subset was required by the fact that Google used federated learning on data from millions of phones in its Android ecosystem.

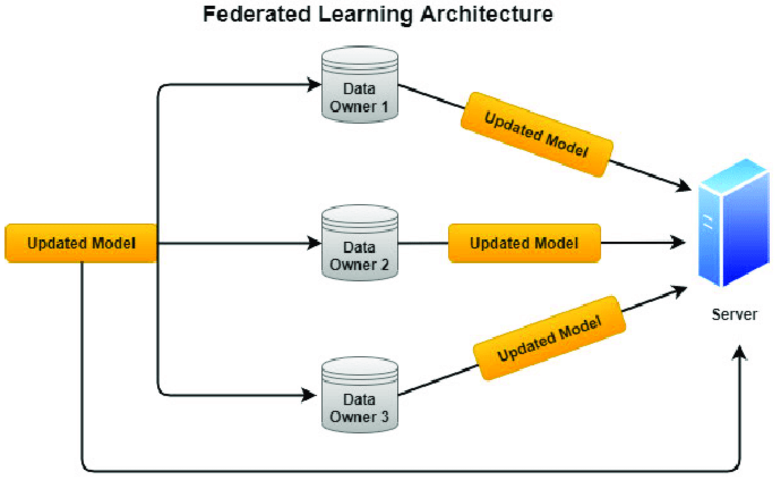


Fig 2- Federated Learning Architecture

A later version of this learning sequence sends gradient updates to the server instead of the weights of the model itself. The general idea is that no original data is ever sent between parties. Only model-related changes are sent. We also know how edge computing helps in this case. So no user data is leaked, only the user gets an initial model that updates overtime on the user side. Then the trained model is sent to the actual server which then finetunes itself from the given models . In this way it is less hassle for the user and maximum security is reached.

Highlighted Factors Found in Recent Years

1. Communication

When building strategies for federated networks, communication is a critical bottleneck to consider. While it is beyond the scope of this paper to present a comprehensive analysis of communication-efficient distributed learning methods, we do highlight a few key trends, which can be divided into three categories: (1) local updating methods, (2) compression schemes, and (3) decentralized training [9].

2. System Heterogeneity

Devices may differ in terms of hardware, network connectivity, and battery power in federated environments, resulting in significant diversity in system parameters across the network. These system features, as shown in Figure 4, make issues like stragglers far more common than in conventional data center setups. We divide various important approaches to dealing with system heterogeneity into three categories: (i) asynchronous communication, (ii) active device sampling, and (ii) fault tolerance[9].

3. Security and Privacy

FL is a promising field of machine learning study. Researchers are putting forth a lot of effort to improve the methodology's ability to deal with privacy and security concerns. For example, the privacy framework given above includes privacy on a local or global level for all network devices. In practice, however, because privacy limitations may differ between devices or even data points on a single device, it may be required to specify privacy on a more granular level. One suggestion is to utilize sample-specific privacy assurances rather than user-specific privacy guarantees, which would provide a less secure kind of privacy in exchange for more accurate models. Developing methods to deal with mixed device- or sample-specific privacy restrictions appears to be a promising direction.

Challenges of Federated Learning

Moving federated learning from a concept to a real-world implementation is not easy. Researchers, even those who don't support federated learning, have helped us better understand the issues that need to be addressed. I think that the most important problems with federated learning are the ones that deal with security, even though there has been a lot of work done on how efficient and accurate it is.

For federated learning, the main reason is to protect the privacy of the data that the clients own. Even if the actual data isn't shown, the weight changes in the model can be used to show properties that aren't the same for everyone who worked on the project. Both the server side and the (other) client side can be used to make this kind of guess. The use of differential privacy is a common way to deal with this risk. Some researchers have looked into the possibility that misbehaving clients could add backdoor functionality or use Sybil attacks to harm the global model. To be able to stop these attacks, more time must be spent on Sybil detection.

Applications of Federated Learning

1.Healthcare

Due to the pandemic situation, we have seen a lot of changes in the last year. During this time, the healthcare industry's shortage of resources was clear.

As a result, healthcare personnel require dependable technology to assist them in providing better care to their patients. However, training an algorithm for clinical use would necessitate a large and diverse data collection.

With strong restrictions like HIPPA in place, sharing crucial information becomes more difficult.

This is where FL entered the picture. On their in-house data pool, participating institutions train the same algorithm.

These trained algorithms could be beneficial to these institutions. It would allow them to gain access to numerous regulations while simultaneously allowing them to work around them. This also gives them access to a larger pool of data from which to learn.

2.Fintech

FinTech refers to businesses that use technology to conduct financial transactions. It serves both individuals and corporations. The words "Finance" and "Technology" are sometimes abbreviated as "FinTech."

The number of data protection legislation is constantly increasing. This allows consumers and organizations to have confidence in one another's ability to keep data safe and secure.

Businesses that rely on FinTech encounter a number of challenges when using traditional machine learning. These concerns include obtaining clearance and lawful consent, data preservation, and the time and expense of gathering and transporting data across networks.

FL offers a straightforward answer in this case. That is, we may employ edge devices and computer resources by keeping the data local.

FL is a distributed and encrypted machine learning method. It enables for cooperative machine learning training on decentralized data without the need for data transmission between participants.

FL has its advantages. It has the ability to resolve and supply FinTech solutions. This is accomplished by looking for data breaches as well as ATO (Account Takeover) Fraud.

It also analyzes credit scores and learns a user's digital footprint to prevent fraudulent actions KYC without having to send data to the cloud.

FL makes it possible for Fintech to mitigate risks. For its customers and enterprises, it develops fresh and inventive techniques. It establishes a foundation of trust between the two parties. It also enables them to develop a more mature relationship.

3.IoT

As technology advances, so does the amount of information available. As a result, there are now additional privacy restrictions in place to protect such information.

Many businesses have begun to use federated learning. They don't share data and instead train their algorithms on a variety of datasets.

Federated learning tries to protect data collected through several channels. It also keeps important information close at hand.

FL is a device-based machine learning solution that does not require the user's personal data to be sent to a central cloud.

As a result, federated learning can aid personalisation. As well as improving device performance in IoT applications.

4. Natural Language Processing

For text prediction, companies like Google[7] employ Federated Averaging techniques in their smartphone keyboard. FL was used to predict the next word on a mobile keyboard. A federated averaging strategy was used to learn a Coupled Input and Forget Gate variant of the LSTM

(CIFG). The FL approach, according to the researchers, can produce better precise recall than server-based log data training.

Companies such as Apple[8], for example, use FL techniques and variants such as Federated Tuning (FT) on their devices to do a combination of on-device processing and suggestions while maintaining customer privacy. Applications centered on FE and FT account for a significant portion of Apple's system consumption. Federated evaluation (FE) is based on the history of user interactions. When compared to live A/B testing, this significantly reduces turnaround times. Before exposing end customers to these candidates via live A/B experimentation, FE can assist swiftly discover the most promising ML system or model candidates.

5. Recommender Systems

When it comes to creating recommendation systems, the federated collaborative filter method is very common. The item-factor matrix is educated in a global server using a stochastic gradient technique by pooling local changes. When compared to the centralized technique, the method is said to have no accuracy loss. A federated matrix factorisation architecture is employed in another way. Federated SGD is used to learn the matrices in this case. Popular recommendation algorithms have been implemented with SMC protocols in the Federated recommender system (FedRecSys). Matrix factorisation, singular value decomposition (SVD), factorisation machine, and deep learning are among the algorithms.

Collaboration between research institutes, hospitals, and federal agencies is essential in modern health systems. Furthermore, in a pandemic situation, international collaboration is critical, but not at the expense of privacy. Because FL ensures secrecy, it allows for collaboration. There is unlikely to be a central server in a healthcare federation. Another difficult aspect is the construction of a decentralized FLS that is also resistant to malefactors. Additional technologies such as secure multi-party computation and differential privacy can address the privacy issue. Explainability of FL models is an open subject, according to a survey on FLS by Li et al[2].

Future Directions

In federated learning, how much communication is required remains to be seen. Indeed, it is generally known that machine learning optimization algorithms can tolerate a lack of precision; in fact, this imperfection can aid generalization. While oneshot or divide-and-conquer communication strategies have been studied in traditional data center environments, their behavior in vast or statistical heterogeneous networks remains unknown. Similarly, one-shot/ few-shot strategies have recently been presented for the federated setting , although they have yet to be conceptually explored or evaluated at scale.

We are at a critical juncture in the development of federated learning, as it is still a new subject. We must ensure that the innovations made in this area are anchored in real-world settings, assumptions, and datasets. It's vital for the larger research community to expand on existing implementations and benchmarking tools, such as LEAF [10] to make empirical results more reproducible and innovative federated learning solutions more widely available.

Contribution

The introduction , History , Basic principles of Federated learning and Architectural overview, Challenges has been written by Shati, S.M Moshiuzzaman (21-92225-3).
Highlighted Factors Found in Recent Years, Applications of Federated Learning and Future Directions has been written by Mujahid Abdullah Al (22-92340-1)

References

[1] Jan Philipp Albrecht. How the gdpr will change the world. Eur. Data Prot. L. Rev., 2:287, 2016.

[2] Q. Li et al., "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3124599.

[3] Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D., & Passerat-Palmbach, J. (2018). A generic framework for privacy preserving deep learning (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1811.04017

[4] Ryffel, T. (2021, November 17). Federated learning in 10 lines. OpenMined Blog. Retrieved April 24, 2022, from https://blog.openmined.org/upgrade-to-federated-learning-in-10-lines/

[5] Pierce, J. (2021, July 14). *Privacy and cybersecurity: A Global Year-End Review*. Inside Privacy. Retrieved April 24, 2022, from https://www.insideprivacy.com/data-privacy/privacy-and-cybersecurity-a-global-year-end-review/

[6] P. Popovski et al. "Wireless access for ultra-reliable low-latency communication: Principles and building blocks" IEEE Netw. vol. 32 no. 2 pp. 16-23 Mar./Apr. 2018. https://www.insideprivacy.com/data-privacy/privacy-and-cybersecurity-a-global-year-end-review/

[7] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Franc¸oise Beaufays Sean Augenstein, Hubert Eichner, Chloe Kiddon, Daniel Ramage - FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION , 2018

[8] How Apple Tuned Up FL - https://analyticsindiamag.com/how-apple-tuned-up-federated-learning-for-its-iphones/

[9] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, Virginia Smith - *Federated Learning: Challenges, Methods, and Future Directions,* August 2019

[10] S. Caldas, P. Wu, T. Li, J. Koneˇcny, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated ` settings. arXiv preprint arXiv:1812.01097, 2018.