Basic of Federated Learning

University of Michigan Dearborn

CIS 695: Master's Project
Summer- 2021

Professor : Dr. Zheng Song

Created by: Shruti Basutkar
Date -5/21/2021

Introduction

Motivation:

- Data is important in the Artificial Intelligence (AI) technology, it is use to train a model to perform various task such as face-recognition, object detection, future prediction.
- **Traditional way:** Centralize learning where data send to central location (server) to train and build Machine Learning (ML) models. There is possibility that owner can lose their control over the data.
- **Concerns**: data privacy and confidentiality
- Hence, regulatory bodies to control data collections and transactions such as General Data Protection Regulation (GDPR) by the European Union (EU) etc.
- Using traditional way and applying privacy laws on data sharing is not possible to share data among different sites in secure and efficient way.

Introduction cont..

Possible Solution:

Federated learning (FL)

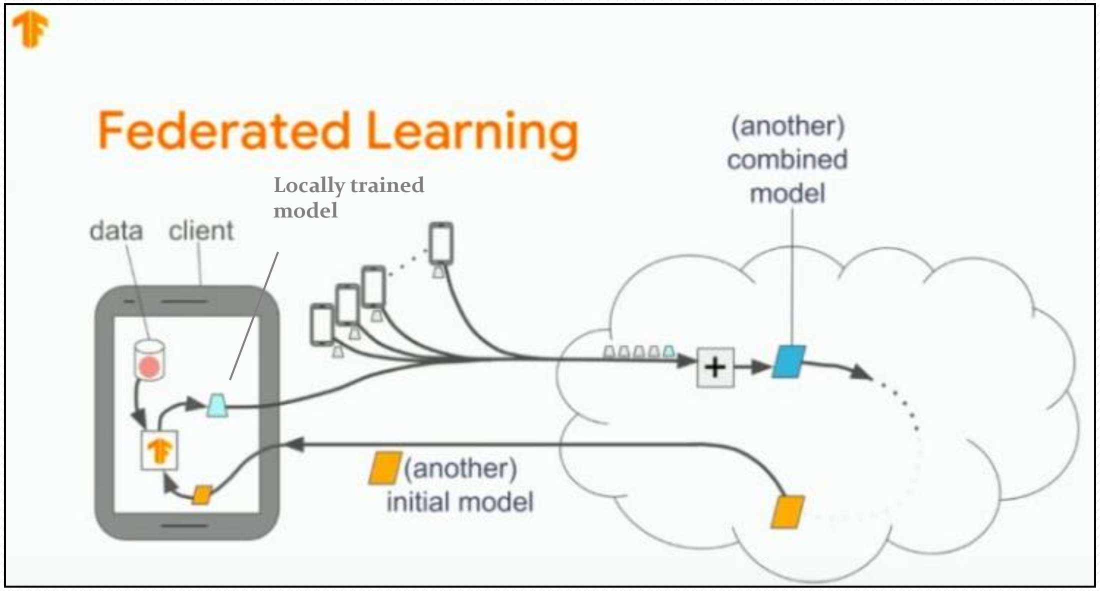
- Idea of decentralize learning
- It is approach to train model at their location where data is resides and then send their respective model updates to sever in order to reach a consensus for a global model.
- Model gets encrypted and shipped to the cloud server
- Enable device to learn from other devices without sharing their data to each other
- Data never leaves the from original owners
- It provides a privacy-preserving mechanism during model learning process.
- **Example**: prediction model in G-board system for auto prediction to completion of words

What is Federated learning (FL)

- It is a Distributed Machine Learning techniques to train model based on data located at multiple sites without sharing data
- Address critical issues such data privacy and data security, data access rights and access to heterogeneous data
- Applications are spread over a number of industries including defense, telecommunications, internet of Things (IoT), and pharmaceutics.

How does FL works

- Devices download generic machine learning model from server
- Personalized and improve its model using its own local data
- Updated models (model parameters weight, bias and other parameters) Send back to the server, training data never leaves from device
- Server aggregates models and obtain a single combined model
- This iterative process until it achieve high-quality model is obtained



https://www.kdnuggets.com/2020/04/federated-learning-introduction.html

Research in Federated Learning:

- Focused on improving security and statistical challenges (secureBoost framework in Vertical FL)
- Effectively adapted to various secure multi-party ML tasks, As a result, a target-domain party can build more flexible and powerful models by leveraging rich labels from a source-domain party.
- A number of strategies to carry out model poisoning attack from malicious data from participant were investigated.
- Re-examining the existing ML (example reinforcement) models under the federated learning settings has become a new research direction in order to protect the privacy of the data and models.
- In field of Computer vision such as in medical image analysis, natural language processing (NLP), recommendation system, Google G-board for next word predictions

Open source project:

Below are few open source projects on research and development in FL algorithms for computation:

- **WeBank Federated AI Technology Enabler**:
- provide secure and computing framework which includes logistic regression, tree-based algorithms, DL and transfer learning

- **TensorFlow Federated:**
- this provides developer existing FL algorithms on their models and data, inaddition to that they can experiment with new algorithms

- **TensorFlow-Encrypted:**
- it provide interface on top of TenserFlow to experiment privacy-preserving ML, no need to be expert in ML, cryptography, distributed system and computing

- **coMind:**
- this is implementation Federated averaging algorithm for training privacy-preserving and data security. It is built on top of TensorFlow

- **Horovod:**
- it works on top of TensorFlow and PyTorchto make distributed DL fast and easy using message passing interface (MPI)

- **OpenMined/PySyft:**
- for building secure and scalable ML models, extension of PyTorch

- **LEAFBeanchmark:**
- includes a suite of open-source federated datasets, rigorous evaluation framework, and a set of reference implementations, aiming to capture the reality, obstacles, and intricacies of practical federated learning environments

Privacy-preserving ML

Privacy-preserving ML gives Background knowledge related to FL , its framework and implementation

Privacy-Preserving Machine Learning (PPML) :

- Techniques for privacy-preserving properties to be built inside machine learning (ML) systems
- Adversaries violets the privacy and confidentiality of ML system

Secure ML:

- Adversaries violate the integrity and availability of ML system

Privacy-preserving ML cont..

Threads and security Model:

1. Privacy thread Model
2. Adversary and secure model

Privacy-preserving ML cont..

1. **Privacy thread Model:**

Attack can be at any stage of ML as follows:

- Attribute inference attacks during publishing stage
- Reconstruction attacks during ML model training
- Inversion or membership-inference attacks during inference phase

Privacy-preserving ML cont..

2. Adversary and secure model:

- **Semi-honest adversaries (honest but curious, passive)**:
- adversaries are abide by the protocol but still it will attempt to learn more information beyond the output

- **Malicious adversaries (active):**
- Malicious behavior break the rules to follow ML protocol honestly

Privacy-preserving ML cont..

Privacy preserving techniques:

FL uses in their frame work different privacy preserving techniques.

Three approaches are:

1. Secure Multi-party computation (MPC)
2. Homomorphic Encryption (HE)
3. Differential privacy (DP)

Privacy-preserving ML cont..

1. **Secure Multi-party computation (MPC): a.k.a Secure Function Evaluation (SEE):**

It compute a function by each party without revealing inputs to the other parties

- **implemented through three different frameworks:**

  Oblivious Transfer (OT), Secret Sharing (SS), Threshold Homomorphic Encryption (THE)

- **Oblivious Transfer (OT):** sender owns a database of message-index pairs (M_i, i)., where 1<i<N. during each transfer receiver choses an index i and to get M_i Model. In this way receiver does not know any information about the database.

- **Secret Sharing:** concept of hiding a secrete value by splitting it into random parts (a.k.a shares) and distribute these to different parties

Privacy-preserving ML cont..

2. Homomorphic Encryption (HE):

- various encryption schema proposed by many researchers. it is solution to perform computation over cipher-text without decrypting the cipher-text .
- It consist four function i.e Key generation, Encryption, Decryption and Evaluation.
- H = {KeyGen, Enc, Dec, Evalg}

Where,

KeyGen: pair of keys {pk, sk} , pk = public key and sk = secret key

Enc: using public key and m input , generates the ciphertext c= Encpk(m) as output

Dec: sk and c as input to retriev plaintext m = Decsk(c)

Eval: c and pk as input and output a ciphertext to a functioned plaintext

Privacy-preserving ML cont..

3. **Differential privacy (DP):**

- To resist the membership inference attack
- Can achieve by adding noise to the data and choosing noise according to an exponential distribution among discrete values

FL classification

Classification according to how data is partitioned among various parties in the feature and sample spaces:

A. Horizontal Federated Learning (HFL)

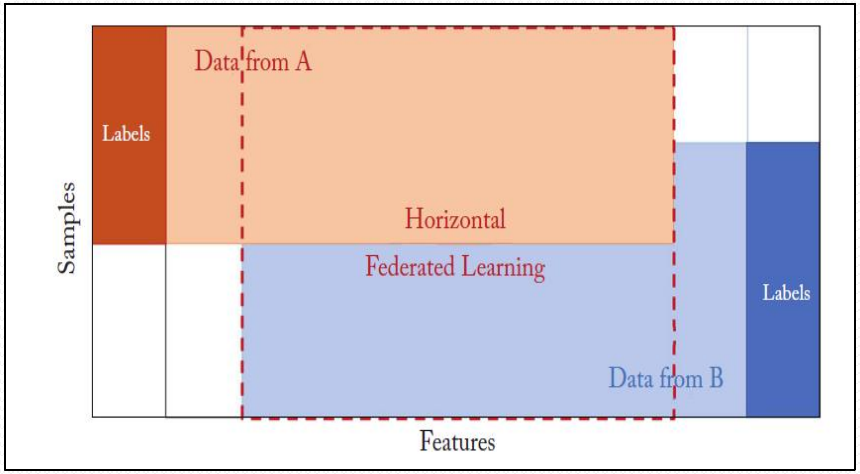
B. Vertical Federated Learning (VFL)

C. Transfer Federated Learning (TFL)

FL classification Cont..

A. Horizontal federated learning (HFL) :

- Participants share overlapping data features differ in data samples
- It resembles the situation that the data is horizontally partitioned inside a tabular view
- For example, two regional banks with different user groups from their respective regions, and the intersection set of their users is very small. However, their business models are very similar. Hence, the feature spaces of their datasets are the same



Framework of Horizontal federated learning (HFL)

FL classification Cont..

- **HFL architecture:**

1. Client-server architecture
2. Peer to peer architecture

FL classification Cont..

1. Client-server architecture

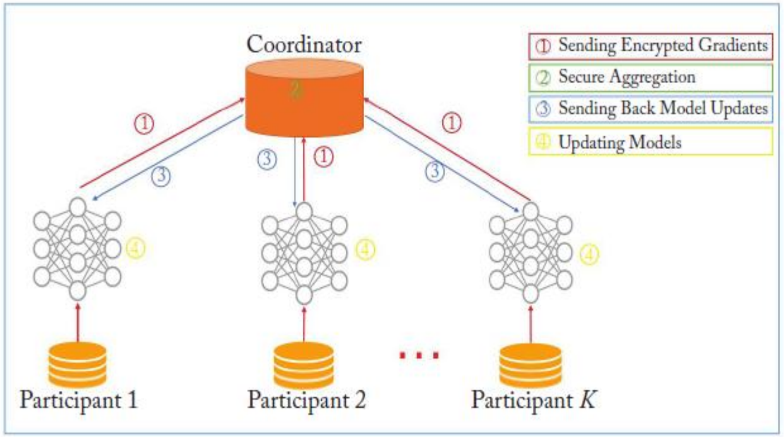
- Where k participants
- Assumptions: participant are honest and server are honest-but-curious, hence chance of information leakage from any participant to server

- **Training process is in four steps**:
- **Step 1:**
- Participants locally compute training gradients, mask a selection of gradients with encryption, differential privacy, or secret sharing techniques, and send the masked results to the server.

- **Step 2:**
- The server performs secure aggregation, e.g., via taking weighted average.

- **Step 3**:
- The server sends back the aggregated results to the participants.

- **Step 4:**
- The participants update their respective models with the decrypted gradients.



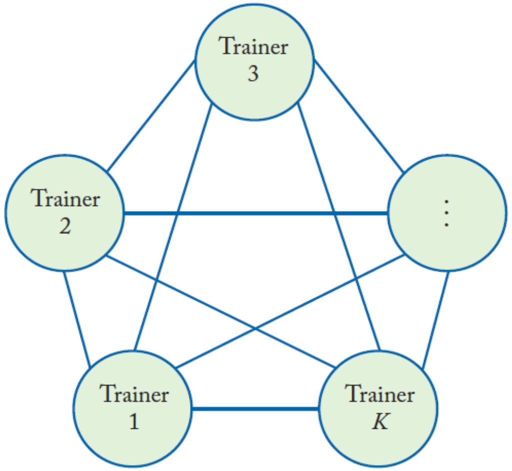
FL classification Cont..

- Above iterative process will continue until loss function converges.
- Participant and server exchanges gradient during above process its call *gradient averaging* a.k.a *synchronous stochastic gradient descent (**SGD**) or federated SGD (**FedSGD**)*
- Instead of gradient if participant and server exchanges model weight that process is called *model averaging*
- *Model and gradient averaging* are referred as *federated averaging (**FedAvg**)*
- Above architecture is capable of preventing data leakage against semi-honest server, *gradient aggregation* is performed with secure *MPC* or *HE*.

FL classification Cont..

2. Peer to peer architecture:

- no central server or coordinator
- K participants a.k.a. trainers or distributed trainers or workers
- Trainer train their model using local data and then using secure channels transfer their model weights to each other.
- public key (PK) based encryption schemes.

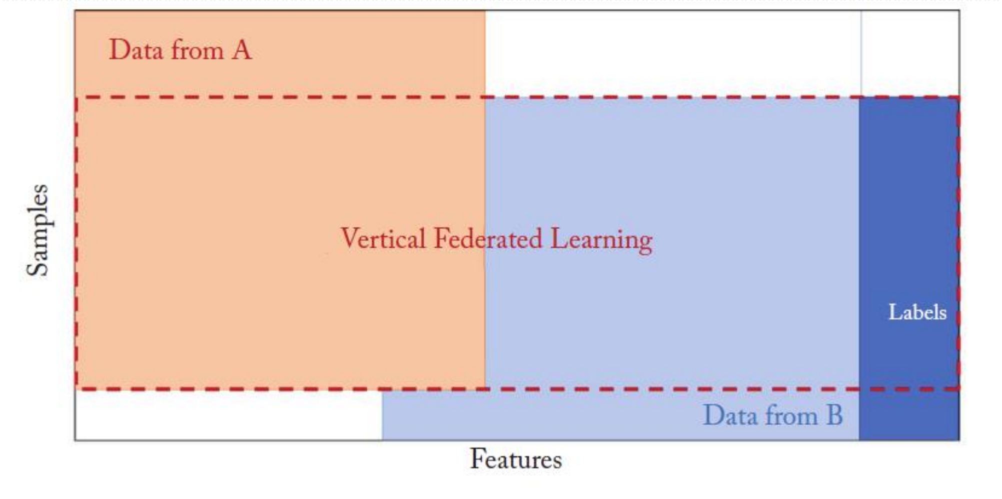


- **There are mainly two ways 1) cyclic 2) Random transfer**
- **Cyclic**: trainers are organized into a chain, first trainer sends model weights to downstream trainer which then updates model using their own dataset and then pass updated model to next downstream trainer.
- **Random**: it follows the *Gossip Learning* method where each trainer decide next random trainer using equal probability trainer to sends its model weights

FL classification Cont..

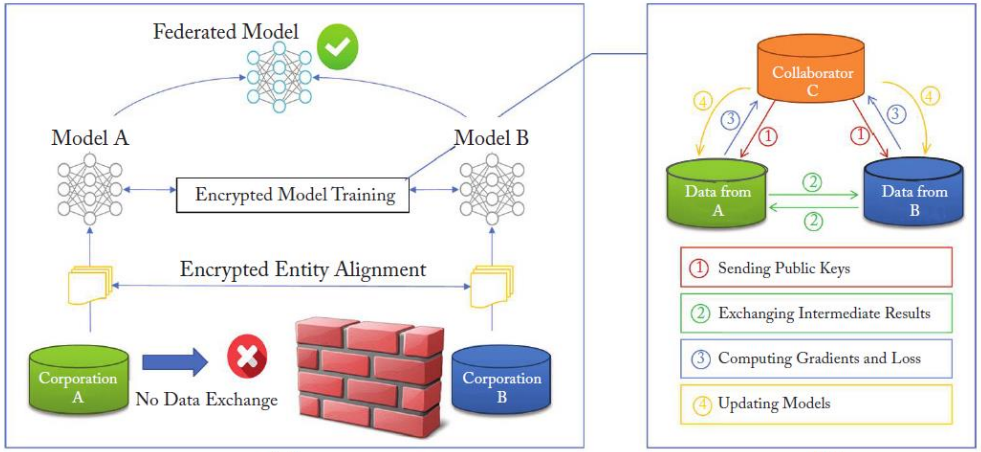
B. Vertical federated learning (VFL) :

- participants share overlapping data samples differ in data features
- It resembles the situation that data is vertically partitioned inside a tabular view.
- Example: Hospitals can collaborate with pharmaceutical companies to make use of the medical records of common patients so as to treat chronic diseases and to reduce risks of future hospitalization.



FL classification Cont..

- **VFL architecture:**
- Suppose companies A and B honest-but curious participant to train an ML model with their own data.
- To ensure privacy and security during training process third party C which is honest participant involved
- Two part process:
- **Part 1**: Encrypted entity alignment:
- uses an encryption-based user ID alignment technique to confirm the common users
- shared by both parties exposing their respective raw data
- **Part 2:**
- Common entices data use to train a joint ML model

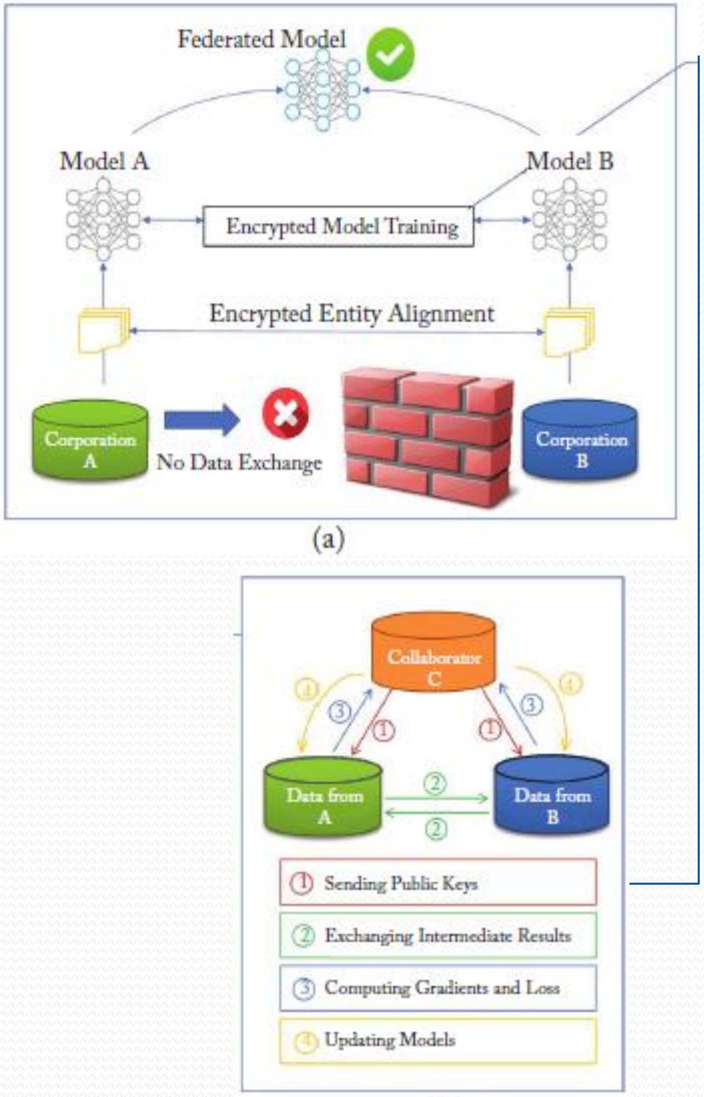


(a) (b)

FL classification Cont.

Training process:

- C created encryption pairs and sends PK to A and B
- A and B encrypt and exchange the intermediate results for gradients and loss calculations
- A and B compute encrypted gradients and add an *additional mask*, respectively. B also computes the encrypted loss. A and B send encrypted results to C
- C decrypts gradients and loss and sends the results back to A and B. A and B unmask the gradients, and update the model parameters accordingly.



(a)

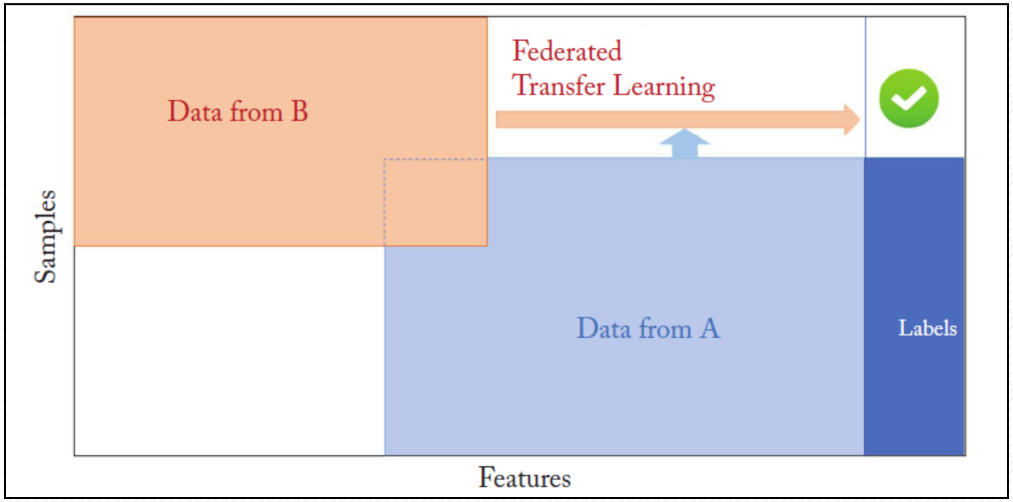
(b)

FL classification Cont..

C. Federated Transfer Learning (FTL) [3]:

- federated learning model combined with transfer learning
- Involved in the case where participant maintain heterogeneous i.e not enough sample or space among participants
- Cross-domain knowledge transfer between source and target parties

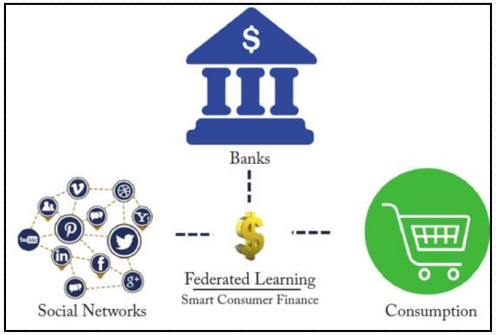
- small overlap in both feature space and sample space
- FTL covers the region in right upper corner by transferring knowledge from non-overlapping features to the new samples
- A predictive model learned from feature representations of aligned samples belonging to party A and party B is utilized to predict labels
- for unlabeled samples of party B.



Applications

Finance:

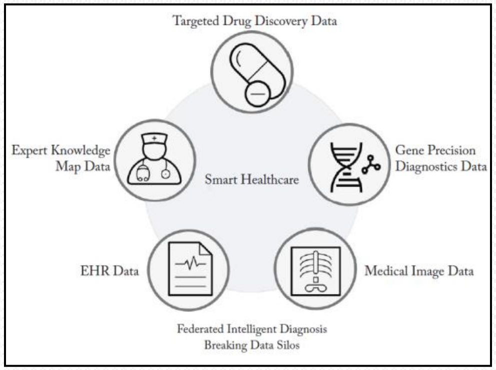
- an build local personalized models without exposing their data
- can apply TFL to address the data heterogeneity problem



Federated learning in smart consumer finance.

Healthcare:

- Medical data are sensitive , so FL allows participants to collaboratively train a shared model without exchanging or exposing their private patient data
- can play an important role in the development of intelligent medical systems



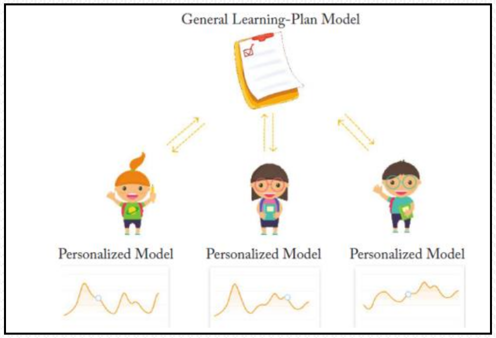
Targeted Drug Discovery Data
Expert Knowledge Map Data
Smart Healthcare
Gene Precision Diagnostics Data
EHR Data
Medical Image Data
Federated Intelligent Diagnosis
Breaking Data Silos

Federated learning in intelligent diagnosis.

Education:

- can help achieve personalized education
- to collaboratively build a general learning-plan model using data stored at students' personal devices
- can provide personalized learning instructions can be built for each student based on that student's strengths, needs, skills, and interests.

More applications such as urban computing and IoT, 5G Mobile networks etc.



Federated learning in education.

Summary

FL is collaborative and decentralized approach to machine Learning with below features

- **Ensure privacy** by providing privacy-preserving mechanism to effectively leverage
- Because the updated model may be utilized to make predictions on the user's device, there is **less latency**.
- Because models are trained on a user's device, there is **less power consumption**.

It is relatively new concept, there is more to come in the future

Research opportunity in ML , statistics, information security , encryption and model compression game theory, indoor localization and economic principles etc.

References

1. https://www.kdnuggets.com/2020/04/federated-learning-introduction.html
2. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
3. https://arxiv.org/pdf/2010.15561.pdf