AI4EOSC Webinar: Introduction to Federated Learning

Basics of Federated Learning: Tips and Tricks

Khadijeh Alibabaei (khadijeh.alibabaei@kit.edu)
Centralized Learning in Machine Learning

- refers to the traditional approach where all data is gathered and stored in a central location to train a machine learning model.

- involves collecting and combining data from multiple sources into a single dataset before training the model.
Centralized Learning in Machine Learning: Challenges

- **Data Flow Management**: Manage the transfer of large volumes of diverse data quickly and accurately across different organizations.
- **Scalability**
- **Communication Overhead**
- **Intense competition within the industry**.
- **Data Privacy**: Ensuring compliance with strict data protection regulations, such as the GDPR\(^1\) and EU AI ACT\(^2\).

1. [https://gdpr-info.eu/](https://gdpr-info.eu/)
2. [https://artificialintelligenceact.eu/the-act/](https://artificialintelligenceact.eu/the-act/)
Federated Learning in Machine Learning

A method that facilitates multiple peers to collaboratively learn a common prediction model by exchanging model weights while keeping the sensitive data on the local devices (Kairouz et al. (2021) and Khan et al. (2023))
Examples of successful applications of FL

Google already used FL in Gboard Android:
When Gboard suggests a query, your phone stores context and interactions locally. Federated Learning uses this to improve Gboard's suggestions.

https://research.google/blog/federated-learning-collaborative-machine-learning-without-centralized-training-data/
More Examples of successful applications of FL

- Apple has employed federated learning to improve Siri’s voice recognition capabilities while maintaining user privacy\(^1\).

- Predicting oxygen requirements for COVID-19 patients in the ER using chest X-rays and health records (Muto, R., et al. (2022)).


Source: Holger R. Roth, et al. (2023)
Categories Federated Learning

Federated Learning can be categorized as (Khan et al. (2023)):

- **Data distribution**
  - **Cross devices**: the model is decentralized across the edge devices and is trained using the local data on each device.
  - **Cross silos**: where the clients are a typically smaller number of organizations, institutions, or other data silos.

- **Architecture**
  - **Centralized Federated Learning**: server coordinates the training
  - **Decentralized Federated Learning**: the communication is peer to peer

- **Learning model**
  - **Horizontal Federated Learning**: each party has the same feature space but different data samples.
  - **Vertical Federated Learning**: datasets of each party share the same samples/users while holding different features (Liu, Y., et al. (2023)).
Workflow in FL: Communication Strategies

- **Scatter and gather**: global model parameters are distributed to client devices for local training; updated parameters are then aggregated.
- **Cyclic Learning** (Chang, K., et al. (2018)): the server selects a subset of clients. Training is done following a predetermined sequential order set by the server.
- **Swarm Learning** (Warnat-Herresthal, S. (2021)): a decentralized subset of FL where orchestration and aggregation is performed by the clients.
Model Aggregation

Model Aggregation in FL is a further development of distributed learning that is specifically tailored to the challenges of unbalanced and non-independent, non-identically distributed data (non-IID).

- **FedAvg**: Local weights are collected and aggregated again after local training, using weighted average.
- **FedProx**: Loss function added to penalize the local weights of clients deviating from the global model.
- **FedOpt**: Added option of using a specified Optimizer and Learning Rate Scheduler when updating the global model (like SGD to aggregate the weights of the model).
- **Scaffold**: Added correction term to the network parameters during local training by calculating the discrepancy between the global parameters.
- **Ditto**: is a method for federated learning that improves fairness and robustness by personalizing the learning objective for each device.
Possible Issues with Federated Learning!

Reconstruction attack (Truong et al. (2021)):
- The original training data samples can be reconstructed from the model weights.

- membership tracing i.e., to check if a given data point belongs to a training dataset, or when a participant whose local data has a certain property, joined collaborative training.


Geiping, J. et.al, (2020)

Inferring that a participant whose local data has the property of interest has joined the training. Melis, L. et.al, (2019)
**Solutions**

- **Data Anonymization**: a technique to hide or remove sensitive attributes, such as personally identifiable information (PII) (Narayanan, A. & Shmatikov, V. (2008)).

- **Differential Privacy (DP)**¹:
  - It provides a formal definition of privacy by introducing noise to query responses to prevent the disclosure of sensitive information.
  - Differential privacy mechanisms include Laplace noise addition, exponential mechanism, and more.

---

1. [https://github.com/google/differential-privacy](https://github.com/google/differential-privacy)
Secure Multi-party Computation (SMPC) (Zapechnikov (2022)): is a cryptographic technique that enables multiple parties to jointly compute a function over their private inputs while keeping those inputs confidential.

- Example: Additive secret sharing

<table>
<thead>
<tr>
<th>Ana</th>
<th>Jorge</th>
<th>Carolin</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>-80</td>
<td>100</td>
<td>180</td>
</tr>
<tr>
<td>0</td>
<td>350</td>
<td>-50</td>
</tr>
</tbody>
</table>
Solutions

- **Secure Model Aggregation (SMA):** in the same way as SMPC, here, the server works with encrypted models in which the individual contributions of the clients remain unknown during the aggregation process.

Secure model aggregation using masked model. Figure from: Lu, S., et.al. (2023).
Solutions

- **Homomorphic Encryption (HE)** (Behera et al. (2020)): allows computations to be performed on encrypted data.
  - **Fully Homomorphic Encryption (FHE)**: allows to perform any number of operations.
  - **Somewhat Homomorphic Encryption (SWHE)**: limits the number of operations that can be performed on encrypted data.
  - **Partially Homomorphic Encryption (PHE)**: allows only one type of operation to be performed.
Another possible attack

Poisoning attacks on Federated Learning (Truong et al. (2021)):

During model training in FL, participants can manipulate the training process by introducing arbitrary updates, potentially poisoning the global model.

Solution?

- Model Anomaly Detection
  (Fung, C., Yoon, C.J., Beschastnikh, I., (2018) and Jagielski, M., et.al. (2018)).
- Not applicable when using secure model aggregation

This problem needs more research
Model FL Frameworks

- **Flower**:  
  - is a flexible, easy-to-use and easily understood open-source FL framework.  
  - It is framework-agnostic meaning that nearly every ML model can be easily migrated to the federated setting.  
  - Well-suited for research and study projects.

- **NVIDIA Federated Learning Application Runtime Environment (NVFlare)**:  
  - NVFlare is a business-ready FL framework by Nvidia.  
  - It supports a variety of models,  
  - NVFlare is framework-agnostic.

1. [https://flower.ai/](https://flower.ai/)  
Model FL Frameworks

- **TensorFlow Federated (TFF)**\(^1\):
  - Developed by Google
  - Specifically designed for compatibility with TensorFlow
  - Integrates smoothly with existing TensorFlow workflows

- **PySyft/PyGrid**\(^2,3\): a Python library for secure Federated Learning
  - Developed by OpenMined
  - Compatibility with popular deep learning frameworks like PyTorch and TensorFlow

1. [https://www.tensorflow.org/federated](https://www.tensorflow.org/federated)
2. [https://blog.openmined.org/tag/pysyft/](https://blog.openmined.org/tag/pysyft/)
3. [https://blog.openmined.org/what-is-pygrid-demo/](https://blog.openmined.org/what-is-pygrid-demo/)
Model FL Frameworks

- **Federated AI Technology Enabler (FATE)**: is an open-source Federated Learning platform developed by WeBank’s AI Group.
  - Business-ready FL frameworks
  - The framework comes with a large number of modules
  - It has a backend for the Deep Learning libraries PyTorch and TensorFlow

1. [https://fate.fedai.org/](https://fate.fedai.org/)
Conclusions

Key Considerations for Federated Learning:

- **Optimizing Client Selection:**
  - Use techniques to improve the response efficiency of end devices.
  - Select customers based on data quality and reliability.

- **Aggregation Algorithm Selection:**
  - Choose the most suitable algorithms for data aggregation.
  - Consider scalability, efficiency, and accuracy of algorithms.

- **Framework Customization:**
  - Tailor framework selection to meet specific task requirements.
Conclusions

Key Considerations for Federated Learning:

- **Security Enhancement:**
  - Implement robust security measures for communication and data sharing.
  - Ensure encryption, authentication, and privacy-preserving techniques.

- **Compliance and Ethical Considerations:**
  - Adhere to data privacy regulations and ethical guidelines.
Conclusions

- **Positive aspects of FL:**
  - Data transferring minimization
  - Build a larger and more diverse dataset
  - Train a more general and global model
  - International collaboration

- **Considerable aspects:**
  - Security issues like model poisoning
  - Biases and Fairness
  - Interpretability
References


References


References


References


