b
UNIVERSITÄT
BERN

Distributed Optimal-Transport Clustering for Malicious User Rejection in Federated-Learning VANETs

Lucas Pacheco, Torsten Braun

October 7, 2022

Institut für Informatik, Universität Bern
Federal University of Pará, Brazil

3rd KuVS Fachgespräch "Machine Learning Networking"

*lucas.pacheco@unibe.ch

Overview

Federated Learning

FL over non-IID data

Vehicular Networks and Connected Vehicles

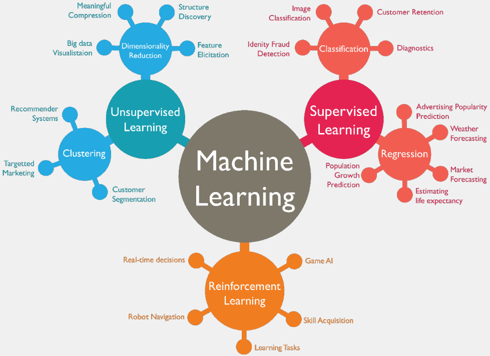
Proposed Algorithm: DOTFL

Conclusions and Future Work

Mobile Edge Computing

- The growth in complexity and usage of ML algorithms has been fueled by the intense data generation of modern devices and computing capabilities.
- Such capabilities are no longer restricted to large providers but contained in user devices at the network's edge.
- **data is generated faster than human specialists can analyze it.**
- **data is generated faster than it can be uploaded for remote processing.**
- **Providers need to craft personalized user experiences for profit**

maximization.



Mobile Edge Computing

- MEC leverages the computing power at the locations where data is generated to process and provide user services.
  - Better context awareness.
  - Increased personalization.
  - Lower latencies and higher throughput at the access network.
- MEC can be deployed at servers near the access network or user devices.

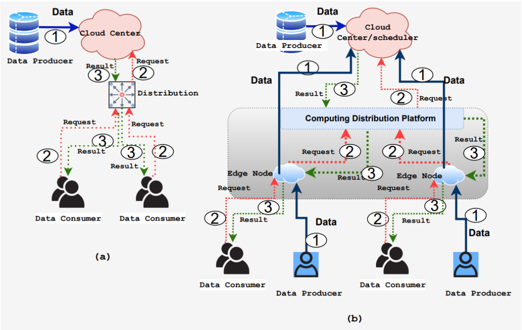


Figure 2. (**a**) Cloud computing paradigm and (**b**) edge computing paradigm.

ML at the Edge

Why and how to perform learning at the edge of the network?

Centralized Training

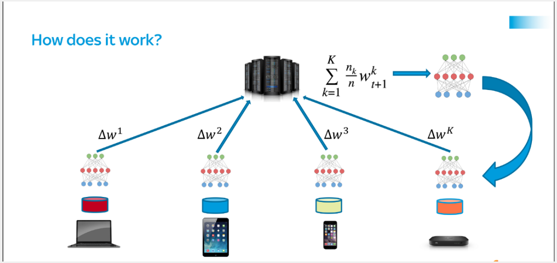
- Training large ML models is a costly operation.
  1. Gather data from users.
  2. Train models at a central location.
  3. Distributed trained models through the network.
- The iterative nature of the ML models training process incurs high costs for network and service providers.

Distributed training.

- Leveraging users' computing capabilities allows for.
  1. Lower training costs.
  2. Higher level of data gathering and aggregation.
  3. More personalized user experiences.

Federated Learning

- Federated Learning, proposed by Google in 2016, pushes the training of AI models to the edge of the network.
- In FL, devices train models on their local datasets and send only trained weights to a central aggregator.
- The biggest advantage of FL is that no raw data samples must be sent through the network.



Federated Learning

- **Strongest point of FL:** high volume and diversity of data.

- **Weakest point of FL:** high volume and diversity of data.

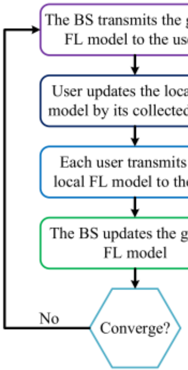
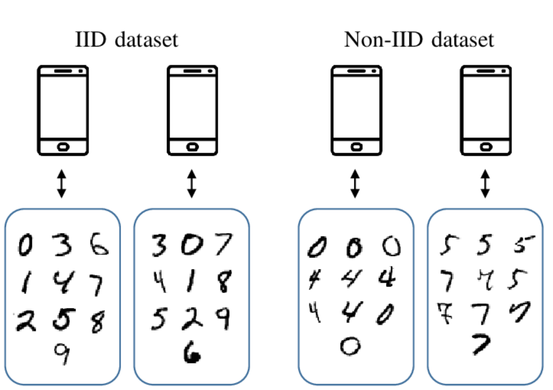

The BS transmits the global FL model to the users
User updates the local FL model by its collected data
Each user transmits the local FL model to the BS
The BS updates the global FL model
Converge?
No

Fig. 2. The learning procedure of an FL algorithm.

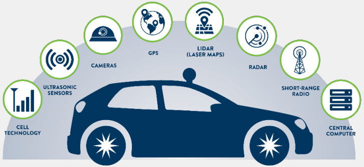
FL over non-IID data



Vehicular Federated Learning

Background and Motivation

- Vehicles are no longer just means of transportation, as they are capable of computing, storage, communication, and data generation.
- The massive amounts of sensitive data generated by sensors and users of vehicles open opportunities for data-driven solutions for transportation and networking.



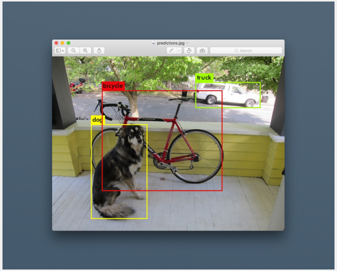
Vehicular Networks and Vehicular Learning

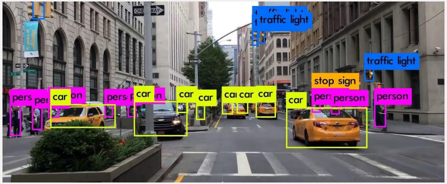
Background and Motivation

- **Vehicles in B5G and 6G networks must offer much more than mobility.**
- Paradigm shift from connected devices to connected, intelligent entities:
  - Object detection.
  - Route planning.
  - Content Delivery.
  - Entertainment.
  - Advertisement.

Vehicular Learning

- Intelligent Transportation Systems.
- Computing capabilities, including Graphical Processing Units.
- Data collection from several sensors for context awareness and decision-making (LiDAR, RADAR, cameras, proximity sensors, etc.).
- Learning tasks for object detection, recognition, tracking, and route planning.
- Objects recognition and image classification are crucial for the correct functioning of CAVs and passenger safety[1].





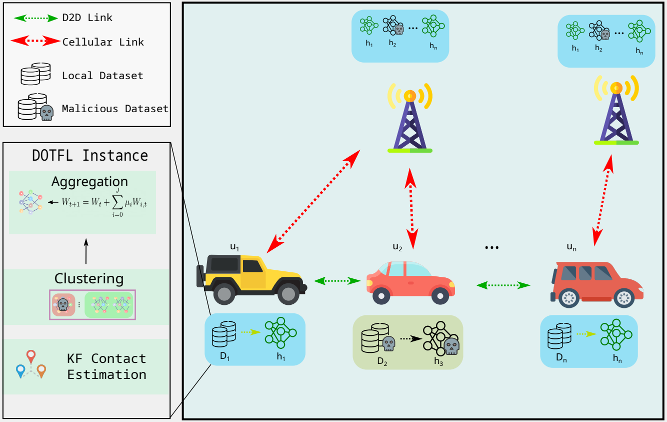
[1] https://www.aljazeera.com/economy/2022/6/15/report-nearly-400-crashes-by-self-driving-cars-in-the-us

Applications and Scenarios

Learning in vehicular scenarios leverages very characteristic opportunities and challenges:

- High mobility: vehicles can be for a very short duration in a given coverage area or contact with neighbor nodes.
- Highly distributed computing and sensing capabilities: vehicles with different computing, storage, and networking capabilities. Various sensors from different manufacturers produce different data features, formats, and biases.
- Possibility for multiple communication interfaces. Vehicles might be able to connect via IEEE802.11p, LTE, 5G, or other interfaces for V2V and V2I. Possibility of integrating pedestrian and parked vehicles in the learning process.

Applications and Scenarios



DOTFL

Distributed OT-based Federated Learning[a]

DOTFL is an Optimal Transport-based Federated Learning algorithm for:

- Training
- Distribution
- and aggregation

Of FL models in vehicular networks.

It is based on the following steps:

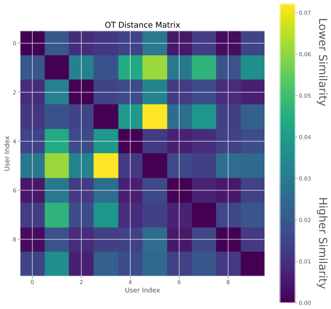
1. Raw model distribution to clients.
2. Local training
3. V2V model distribution
4. Model clustering
5. Aggregation

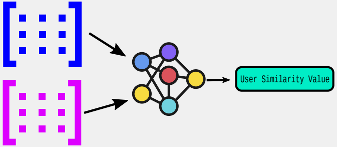
[a] Under review Elsevier Pervasive and Mobile Computing

Federated Clustering

Computing the similarity between users with no access to their data.

- Optimal Transport as a generalized dataset distance metric:
- **Learning OT distance from sample datasets in the cloud.**

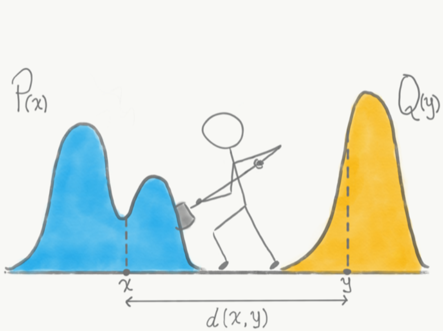




$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p} \tag{1}$$

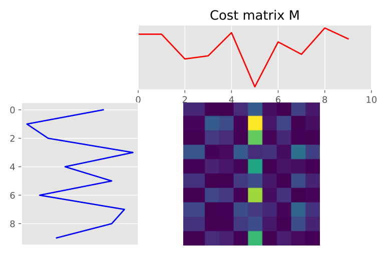
Federated Clustering

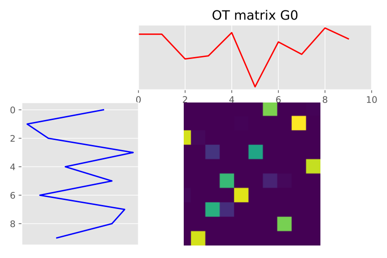
Optimal Transport and Wasserstein distance:

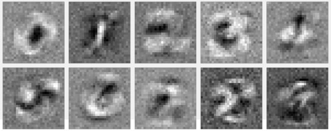


Users' Data Distribution

- Data samples artificially made non-IID.
- Horizontal FL: all users have the same data classes and features.







Semi-decentralized Federated Learning

Continuous Learning at Aggregation Level

Features in the network may change over time, how does the FL instance react to changes?

Alternative Aggregation Functions for Continuous Learning:

- Vehicles keep collecting new data constantly.
- Features may change in the datasets (changes in an urban environment will introduce new knowledge in the network).

Exponential smoothing-based aggregation, given N contributions:

$$W_i = \left(\frac{1-a}{1-a^n} a^{(i-1)} \right), 1 \leq n \leq N \tag{2}$$

Experimental Setup[a]

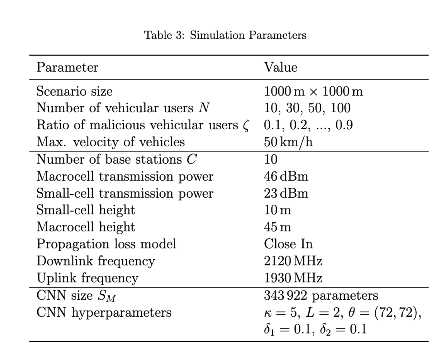


Table 3: Simulation Parameters

Parameter	Value
Scenario size	1000 m × 1000 m
Number of vehicular users N	10, 30, 50, 100
Ratio of malicious vehicular users ζ	0.1, 0.2, ..., 0.9
Max. velocity of vehicles	50 km/h
Number of base stations C	10
Macrocell transmission power	46 dBm
Small-cell transmission power	23 dBm
Small-cell height	10 m
Macrocell height	45 m
Propagation loss model	Close In
Downlink frequency	2120 MHz
Uplink frequency	1930 MHz
CNN size S_M	343 922 parameters
CNN hyperparameters	$\kappa = 5,\ L = 2,\ \theta = (72, 72),$ $\delta_1 = 0.1,\ \delta_2 = 0.1$

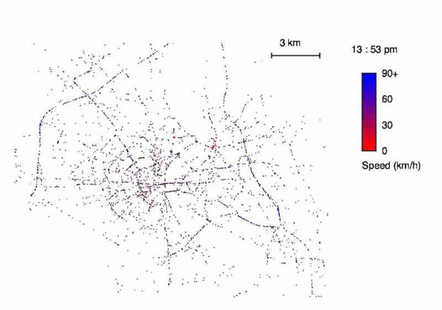


Figure: Simulation Parameters, scenario map

[a] github.com/lsiddd/federated_sid

Experimental Results

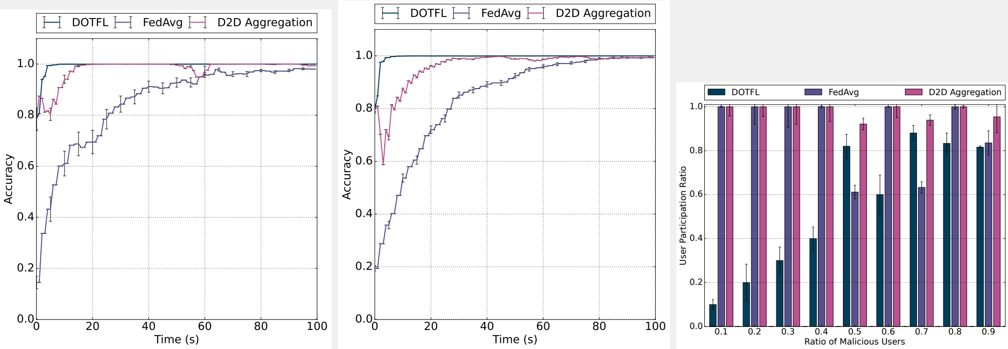


Figure: DOTFL in scenarios with 10 and 100 vehicles and the presence of malicious users.

Shortcomings and Current Challenges

- Lack of a standardized simulation environment for vehicular federated learning.
  - The implemented FEDSID framework supports the simulation of mobility traces (Köln and Luxembourg integrated) and learning via the Keras/TensorFlow modules.
  - Communication simulation can be simplistic (Calculations based on Shannon's law and CI propagation). No support for bit error or probing CPU/GPU usage for devices.
- Individual contributions can be vulnerable to inference attacks. Offset when more models are available in the network.
- High communication volume to share individual contributions. Tests with compression, gradient sharing, and splitting of the model layers are in progress.

Conclusions and Future Work

- The presence of MEC can significantly improve ML tasks on which vehicles depend more on their functioning.
- The presence of MEC capabilities can also be extended to vehicular devices, providing strict requirements for ground users at a low deployment cost.
- Vehicles are equipped with sensors that generate large amounts of data to be processed and trained ML models. Balancing cooperative learning and privacy is an important issue in modern vehicular networks.
- We propose an FL-based algorithm to train personalized, high-quality ML models for vehicles, achieving better learning performance than state-of-the-art algorithms.

References

1. Pacheco, Lucas, et al. "Federated User Clustering for non-IID Federated Learning." Electronic Communications of the EASST 80 (2021).
2. **Under Review** Distributed Optimal-Transport Clustering for Malicious User Rejection in Federated-Learning VANETs. Elsevier Pervasive and Mobile Computing.

Thank you for your attention!

Questions?

`lucas.pacheco@unibe.ch`