A Benchmark of Concept Shift Impact on Federated Learning Models

Comparing the differences in performance between federated and centralized models under concept shift

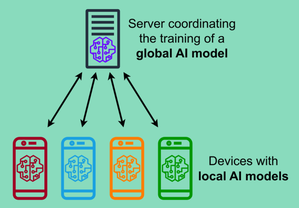
Author: Matei Ivan Tudor (M.T.Ivan-1@student.tudelft.nl)

TU Delft

Supervisor: Swier J.F. Garst (S.J.F.Garst@tudelft.nl)
Responsible Professor: David M.J. Tax (D.M.J.Tax@tudelft.nl)

Introduction to Federated Learning and Concept Shift

- **Federated learning** is a decentralized approach to machine learning, where data is split throughout multiple clients and where privacy is paramount. It enables multiple devices to collaboratively train a model while keeping the data localized.



- **Concept drift** is the phenomenon of data having a shift in its underlying distribution over time, necessitating continuous model adaptation to avoid performance drops.
- We introduce the notion of **concept shift**, where a model is trained on a set of data and deployed in an environment where the data faces a sudden change of concept.

Research Question

Are federated models affected by concept shift more than centralized models?

Problem description and Methodology

- Given a probability distribution P(**x**, y), where **x** is the feature vector and y is its class label, a **concept shift** takes place whenever there is a change in the joint probability distribution P(**x**, y) = P(**x**)P(y|**x**) = P(y)P(**x**|y). A shift can take place in any of the four presented probability distributions.
- This research considers shift in P(**x**), also known as **virtual domain shift,** explored with image data, and in P(y|**x**), also known as **real shift,** explored with tabular data. A **real shift** alters the classification boundary, unlike a **virtual shift**, which does not [1].
- **Concept shift** is defined by two characteristics: form (the probability distribution affected by the shift) and severity (describing how different the new concept is from the one during training).

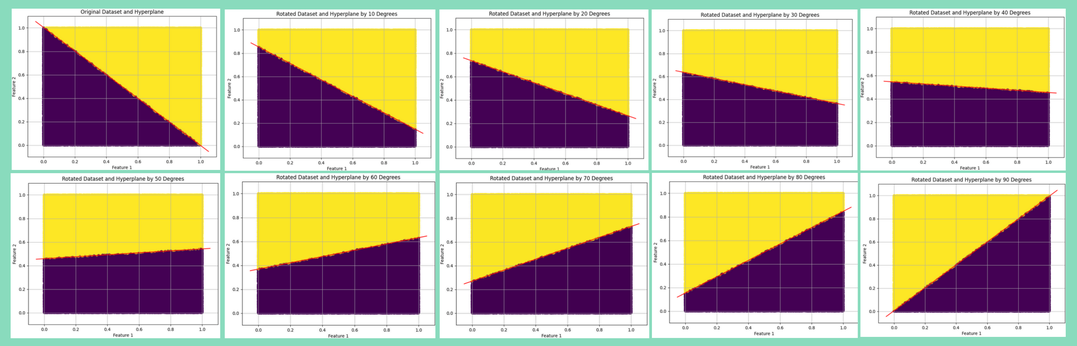
- The classification performance of 3 models is compared: a centralized model, a federated model with IID data throughout clients and a federated model with non-IID data throughout clients, i.e., class sample size differs between the clients.
- A basic FL algorithm is used for training the federated models, FedAvg [2], proven to have similar performance to centralized learning techniques.

References

[1] G. Yang, X. Chen, T. Zhang, S. Wang and Y. Yang, "An Impact Study of Concept Drift in Federated Learning," 2023 IEEE International Conference on Data Mining (ICDM), Shanghai, China, 2023, pp. 1457-1462, doi: 10.1109/ICDM58522.2023.00191.
[2] M. H. Brendan, E. Moore, D. Ramage, S. Hampson, and Arcas, Blaise Agüera y, "Communication-Efficient Learning of Deep Networks from Decentralized Data," arXiv.org, 2016. https://arxiv.org/abs/1602.05629

Inducing Concept Shift

- For tabular data, a 2D linearly separable, binary-classification problem is used.
- **Real shift** is induced by rotating the boundary around its center. Rotation is done counterclockwise incrementally by 10° until 90°. A greater rotation induces a more severe shift.



- The CIFAR-10 dataset is used for image data. **Virtual domain shift** is induced by applying compositions of image transforms to the test set. The intended effect is to induce **concept shift**, rather than to perform data augmentation.
- To measure dissimilarity, the Jensen-Shannon divergence between feature distributions of transformed and original data is used, along with cumulative variance explained by the principal components (for the first n features, denoted as C n). The assumption is that the transform producing most similar data is on augment level ("Train"), while the "Aggressive" transforms represent concept shift. A shift increases in severity when the dissimilarity increases as well.

Original Image

Augment level transform ("Train")

Shift-inducing transform ("Aggressive")

Transforms	JS Divergence	C 1	C 2	C 3	C 10	C 100
None	0	0.29	0.40	0.47	0.65	0.90
Train	0.06	0.42	0.52	0.58	0.73	0.92
Aggressive	0.14	0.65	0.71	0.75	0.84	0.95
Aggressive & Blur	0.17	0.70	0.76	0.80	0.88	0.98
Aggressive & Noise	0.19	0.60	0.66	0.70	0.78	0.91

Differences in the transformed data compared to the original data

Setup and Results

Tabular Data

- Federated training is done with 20% clients per round, solely on non-rotated data.
- The problem is explored with 30, 100, and 200 training samples, tested on 200k samples.
- The problem's complexity is increased by adding two, respectively five redundant features. They do not affect the real boundary of the problem.

- Results show that the non-IID model is underperforming, however given enough training samples, it can achieve comparable performance to IID and centralized model. As the severity of the shift increases, the gap in performance decreases.
- The IID model shows comparable performance to the centralized model, except when training on 30 samples and five noisy features.

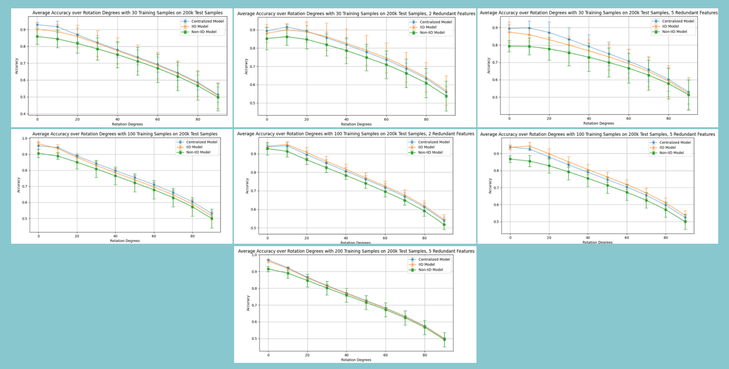


Image Data

- Federated training is done with 100% clients per round. With 20% clients per round, the non-IID model cannot keep up with the other two. The IID model, however, does not face this issue.
- Furthermore, training is done with only original train data, and with 20% of data "Train" transformed, 80% original train data. This is done as the "Train" composition provides features in common with the "Aggressive" transforms, and to understand how well the models can make use of these common features under concept shift. Testing is done on the fully transformed test set.

- Results show that in all five experiments, the centralized model is better. With 100% of the clients, the non-IID and IID models show very similar performances. The impact of decentralized data is still observed, which is due to FedAvg being less precise in weight calculations than the centralized model. Here as well as the severity of the shift increases, the gap in performance decreases.
- Furthermore, results with 20% clients show that data being non-IID is the more significant issue, rather than the decentralized nature of data..

Model, Transformed Data	No Tr.	Train Tr.	Aggressive Tr.	Aggressive & Blur Tr.	Aggressive & Noise Tr.
Centralized, 0%	**84.02%**	69.94%	44.57%	28.58%	18.94%
Centralized, 20%	83.80%	**72.79%**	**51.00%**	**30.16%**	**19.40%**
IID Federated, 0%	82.20%	67.49%	42.89%	26.47%	17.16%
IID Federated, 20%	82.05%	69.61%	46.74%	29.09%	16.94%
n-IID Federated, 0%	79.81%	66.05%	43.18%	26.51%	16.30%
n-IID Federated, 20%	81.33%	69.02%	45.31%	28.39%	16.76%

Classification performance of the three models on test data under the different transforms, with federated models using all clients per round to train

Model, Transformed Data	No Tr.	Train Tr.	Aggressive Tr.	Aggressive & Blur Tr.	Aggressive & Noise Tr.
IID Federated, 0%	81.70%	66.50%	41.76%	25.38%	16.55%
IID Federated, 20%	81.59%	67.43%	44.25%	26.96%	16.68%
n-IID Federated, 0%	74.24%	57.83%	34.51%	21.85%	15.14%
n-IID Federated, 20%	73.81%	58.91%	37.88%	22.61%	15.38%

Classification performance of the federated models on test data under the different transforms, with federated models using 20% clients per round to train

Conclusions

Federated models can indeed be more affected by **concept shift** than centralized models. However, the extent to this difference in performance is also affected by other factors, such as:

- non-IID data throughout the federated clients;
- complexity of the problem and number of available training samples;
- severity of the shift;
- the number of clients used for training.

Experiments have shown that the centralized model is better under **concept shift** than both federated models, and that the IID federated model is better than the non-IID one. However, the overall difference in performance was not large, and might not outweigh the privacy trade-offs of centralizing data.

Future Work

Further work is comprised of experimentation with different types of data (e.g. textual), different problems (e.g., image segmentation), and other datasets to use. Another interesting investigation would be to experiment with a problem with class overlap, as this makes the learning process fundamentally harder.

In the context of the tabular experiment, further work includes experimentation with multi-classification, rather than just binary classification, as well as using a dataset with a non-linear boundary, as the classification problem is more complex in these settings.

Lastly, experimentation should also be done with **concept shift** affecting P(y) and P(**x**|y), as these cases were not considered.