Empowering Pandemic Response with Federated Learning for Protein Sequence Data Analysis

Prakash Chourasia,
Zahra Tayebi,
Sarwan Ali, and
Murray Patterson

Georgia State University
August 1, 2023

Table of Contents

1. Introduction
2. Motivation
3. Methodology
4. Architecture
5. Workflow
6. Dataset
7. Baseline Models
8. Results
9. Conclusion and Future Work

Introduction

Sequence data analysis :

- Studies of Alterations in the protein sequence to classify and predict amino acid changes in SARS-CoV-2 are crucial in
  - Understanding the immune invasion and host-to-host transmission properties of SARS-CoV-2 and its variants
  - Identify transmission patterns of each variant may help policy makers to prevent rapid spread
  - May help in vaccine design and efficacy
- Unravel the mysteries of genetic info & its functional implications
- Phylogenetic tree construction based methods - a Traditional way to trace evolution.
- Later Machine Learning and Deep Learning played major role.

Introduction

Machine Learning and Deep Learniing : Several work is done while using *k*-mers and a kernel-based approach to classifying the spike sequences.

- Not memory efficient, not scalable, data privacy concerns.
- Deep learning - several SOTA approaches for medical image data
- Are centralized system based approach

Federated Learning :

- It is a ML paradigm to decentralize the processing and model training.
- The concept of taking the model to the client instead of taking data to the model.
- Protect the data privacy, personalized models are possible.
- Reduced computational cost and latency issue.

Motivation

- Passive participation of Countries during COVID-19
- Enable real-time surveillance of epidemics - encourages countries to contribute factual and legit data
- Privacy concerns in biological sequences.
- Reduced computational cost.
- Personalized FL model to address local concerns

- Genomic surveillance: Tracking the spread of pathogens
- Real time identification of new and rapidly emerging variants

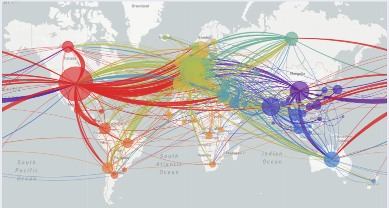


Image Source: `https://genome.ucsc.edu/covid19.html`

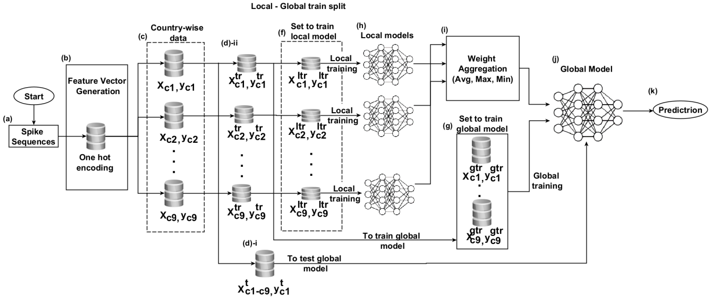
Methodology

- Federated learning-based solution to tackle the issue of data privacy
- Our federated learning-based approach trains the feed-forward neural network locally on local data (hidden from the outside world)
- And aggregates the learning from these local models by taking avg, min, and max weight values from the local models.
- Only updated weights are pushed to the central server (only useful information to detect patterns and not actual data)
- A separate feed-forward neural network is then trained (as a global model) on the aggregated weights and an independent dataset, allowing for the final prediction of coronavirus lineage classification.

Architecture

- Given the input data (A), first compute the One Hot Encoding (OHE) [1] from input spike protein sequences
- The architecture is comprised of two building blocks:
  - Local models and
  - Global model.
- The idea is to distribute the computational load to local clients and extract useful information from the local dataset owned/produced by local clients (countries).
- Consist of feed-forward neural networks both for local and global setting
- We train multiple local models on different country-wise local datasets and then their weights are aggregated to form a global model

Workflow


Local - Global train split
(a)
Start
Spike Sequences
(b)
Feature Vector Generation
One hot encoding
(c)
Country-wise data
X_{c1}, y_{c1}
X_{c2}, y_{c2}
X_{c9}, y_{c9}
(d)-ii
X^{tr}_{c1}, y^{tr}_{c1}
X^{tr}_{c2}, y^{tr}_{c2}
X^{tr}_{c9}, y^{tr}_{c9}
(d)-i
$X^{t}_{c1\text{-}c9}, y^{t}_{c1}$
To test global model
(f)
Set to train local model
$X^{ltr}_{c1}, y^{ltr}_{c1}$
$X^{ltr}_{c2}, y^{ltr}_{c2}$
$X^{ltr}_{c9}, y^{ltr}_{c9}$
Local training
Local training
Local training
To train global model
(h)
Local models
(i)
Weight Aggregation (Avg, Max, Min)
(g)
Set to train global model
$X^{gtr}_{c1}, y^{gtr}_{c1}$
$X^{gtr}_{c9}, y^{gtr}_{c9}$
Global training
(j)
Global Model
(k)
Predictrion

- We initialize 9 local models (each for a separate country in our data)
- Weights from different local models are aggregated using different strategies like *Average* [2], *Minimum*, or *Maximum weight*.
- The global model is trained using the aggregated weights as its initial weights and global training dataset.

Workflow

- Optimization problem, we have to minimize the aggregated loss :

$$L(F) = \frac{1}{k} \sum_{i=1}^{k} L(F, D_i) \tag{1}$$

where $L(F, D_i)$ is the loss on local dataset D_i when using global model F.

- We randomly split each country dataset into 70-30 percent for training sets (X_{ci}^{tr}) and test sets (X_{ci}^{t})
- The test set (X_{ci}^{t}) is reserved (kept unseen) used on final global model
- X_{ci}^{tr} is further divided into 70-30 percent random split as X_{ci}^{ltr} and X_{ci}^{gtr}
- After training local models we aggregate to initialize global model
- There are 3 different aggregation functions (min, max, average) that we are using in our experiments.
- Predictions produced by testing trained global model using test set

Dataset - Country wise distribution

- The Spike7k dataset contains 7000 spike sequences of the SARS-CoV-2 virus that were taken from the well-known database GISAID [a].
- Country-wise distribution for 22 coronavirus variants.
- Each country has several variants in their sequences.

Country	Sequences
USA	1779
England	1662
Germany	470
Denmark	374
Sweden	241
Japan	222
Scotland	194
Canada	191
France	174
Others	1693
Total	7000

[a] https://www.gisaid.org/

Dataset Lineage Distribution

- Dataset statistics for 22 coronavirus variants (total 7000 sequences).
- These are lineage distribution spread-ed across all countries

Lineage	Sequences	Lineage	Sequences
B.1.1.7	3369	B.1.160	92
B.1.617.2	875	B.1.351	81
AY.4	593	B.1.427	65
B.1.2	333	B.1.1.214	64
B.1	292	B.1.1.519	56
B.1.177	243	D.2	55
P.1	194	B.1.221	52
B.1.1	163	B.1.177.21	47
B.1.429	107	B.1.258	46
B.1.526	104	B.1.243	36
AY.12	101	R.1	32
Total	7000	-	-

Baseline Models

- Long Short-Term Memory (LSTM)
- Gated Recurrent Unit (GRU)
- Convolutional Neural Network
- Feed Forward Neural Network
- Poincaré Embedding
- Autoencoder + Neural Tangent Kernel

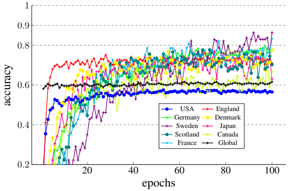
To assess the quality of classification we employ metrics such as average accuracy, precision, recall, weighted F_1, macro F_1, ROC-AUC, and training run-time.

Results

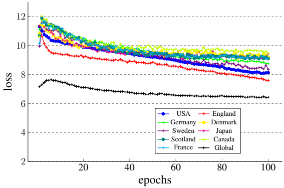
Category	Method	Classifier	Acc. ↑	Prec. ↑	Recall ↑	F1 (Weig.) ↑	F1 (Macro) ↑	ROC AUC ↑	Train Time (sec.) ↓
Feature Engineering	Poincaré embedding	SVM	0.40	0.24	0.40	0.30	0.03	0.49	1248.18
		NB	0.47	0.30	0.47	0.32	0.03	0.50	9.65
		MLP	0.19	0.26	0.19	0.21	0.03	0.49	92.73
		KNN	0.44	0.29	0.44	0.34	0.04	0.50	1.96
		RF	0.48	0.23	0.48	0.31	0.02	0.50	91.68
		LR	0.24	0.25	0.24	0.25	0.04	0.49	1935.23
		DT	0.42	0.43	0.42	0.42	0.11	0.54	200.11
Central NN Baselines	Autoencoder + NTK	SVM	0.48	0.50	0.48	0.47	0.14	0.68	0.011
		NB	0.50	0.46	0.50	0.45	0.22	0.65	0.002
		MLP	0.46	0.46	0.46	0.44	0.19	0.65	0.917
		KNN	0.41	0.33	0.41	0.36	0.08	0.58	0.002
		RF	0.52	0.49	0.52	0.48	**0.19**	0.68	0.185
		LR	0.50	0.48	0.50	0.48	0.17	0.65	0.009
		DT	0.52	**0.54**	0.52	0.52	0.18	**0.69**	**0.001**
	LSTM	-	0.47	0.22	0.47	0.30	0.02	0.50	29872.920
	GRU	-	0.49	0.24	0.49	0.33	0.03	0.50	16191.921
	CNN	-	0.13	0.02	0.07	0.04	0.02	0.49	2902.425
	Feed Forward NN	-	0.62	0.52	0.62	0.55	0.08	0.55	2360.09
Federated Learning	FLAvgWeight (ours)	-	**0.63**	0.53	**0.63**	**0.57**	0.13	0.58	2617.56
	FLMinWeight (ours)	-	0.49	0.24	0.49	0.32	0.03	0.50	2169.74
	FLMaxWeight (ours)	-	0.49	0.24	0.49	0.32	0.03	0.50	2254.53

- We can observe that for all evaluation metrics except the training run-time, the FL-based model with *Avg* aggregation function outperforms all the baselines.
- Note - Data transfer to central server is not accounted in centralised models.

Results



(a) Global Accuracy



(b) Global Loss

- The black line represents the Global model in accuracy and loss.
- The accuracy for local models Figure-a are better.
- Expected because, we definitely want our local model to perform better on the local dataset.
- But at the same time, the model should learn from other countries datasets to be more robust and perform well on new data.
- Similarly, we can see in Figure-b the loss is significantly better in the aggregated Global model.

Conclusion and Future Work

Conclusion

- We demonstarted with experiments that proposed FL based model effective, preserve data privacy, and outperforms traditional centralized deep learning models.
- The proposed approach can prepare us to handle pandemics better

Future Work

- We plan to explore other deep learning models such as recurrent neural networks (RNNs)
- Evaluate the proposed approach in the other domains
- Develop methods to increase data utilization fairness

Thank You

Questions!!

- K. Kuzmin *et al.*, "Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone," *Biochemical and Biophysical Research Communications*, vol. 533, no. 3, pp. 553–558, 2020.

- B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, 2017, pp. 1273–1282.