IEEE International Conference on Computer Communications

(INFOCOM 2021)

Dual Attention-Based Federated Learning for Wireless Traffic Prediction

Chuanting Zhang, Shuping Dang, Basem Shihada, Mohamed-Slim Alouini

King Abdullah University of Science and Technology (KAUST)

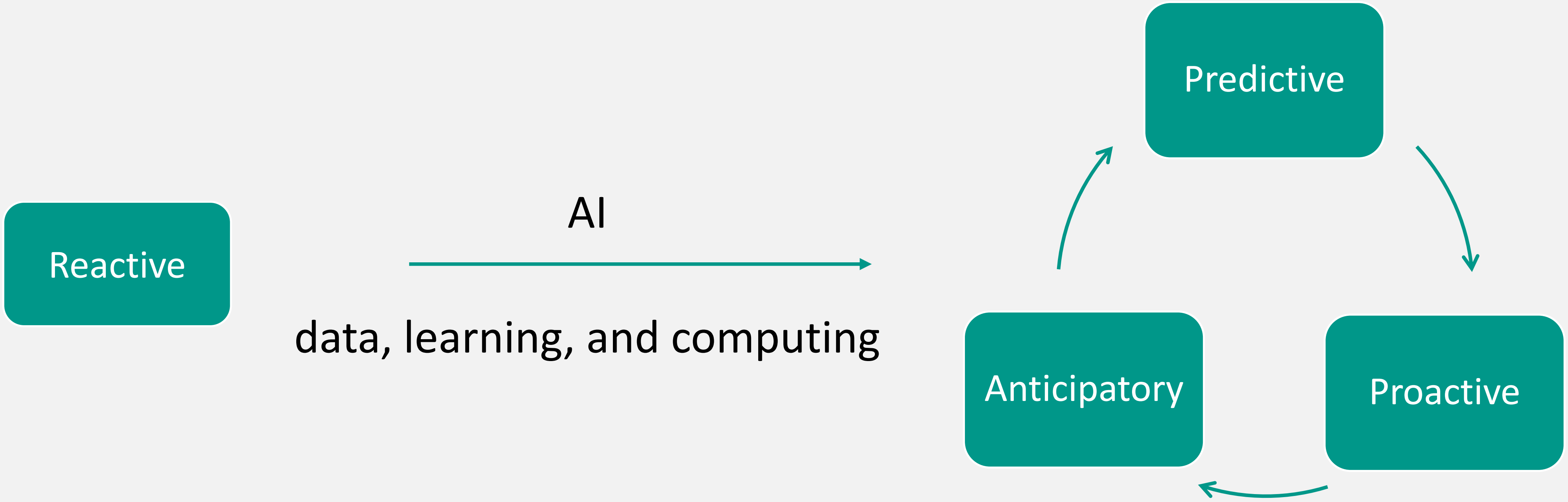
Saudi Arabia

Outline

- Background and Motivation
- Preliminaries and Problem Formulation
- Proposed Method: FedDA
  - Data Augmentation
  - Iterative Clustering
  - Dual Attention-based Model Aggregation
- Evaluation
  - Experimental Settings and Performance Comparisons
  - Parameter Sensitivity
- Summary

Background

- The future networks will be AI-empowered systems
  - Communication systems need AI technologies to make themselves smart enough that can learn and make decisions by themselves.



Background

- Wireless traffic prediction is crucial in future learning-based communication systems, with prediction we can:
  - Improve network management through dynamic congestion control
  - Reduce operating expenditure by accurate radio resource purchase
  - Enhance energy efficiency by intelligent BS on/off

Current Methods and Drawbacks

- Centralized methods, e.g., *ST-DenseNet* and *STC-Net*
  - Need to ***transfer raw data to datacenter*** to learn a generalized model
  - Consume lots of ***bandwidth***
  - May have ***high latency*** for mission-critical tasks
  - Involve ***no cooperation*** from multiple MNO due to data privacy
- Fully distributed methods, e.g., *Gaussian Process Regression*
  - ***Could not capture spatial dependences*** among different BSs/cells/regions
  - May have ***limited data***, especially in places with newly deployed infrastructures
  - Involve ***no cooperation also*** due to data privacy

What Do We Need for Wireless Traffic Prediction

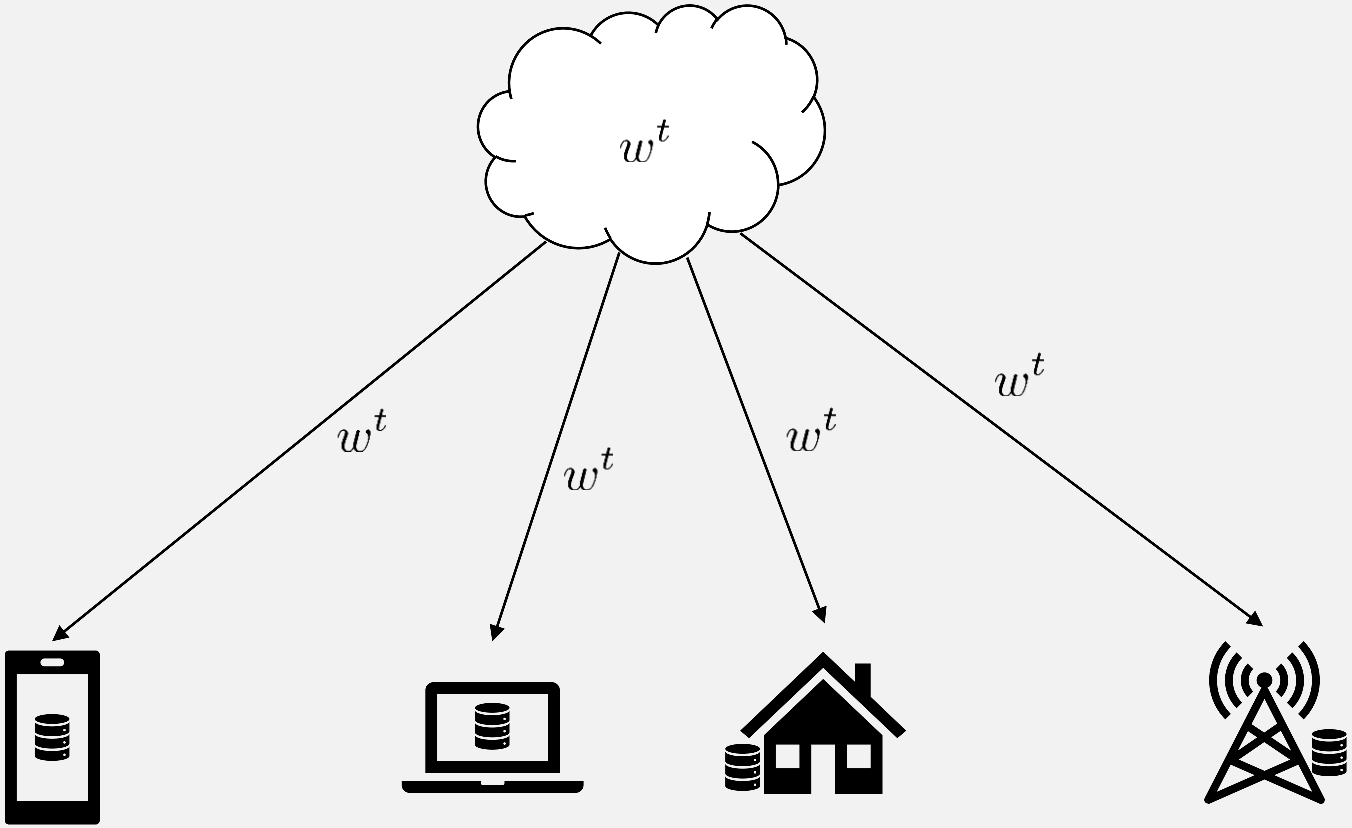
- We need a model that can
  - Capture ***both spatial and temporal dependencies***
  - Be ***deployed at the edge*** to reduce latency
  - ***Without transferring data*** from local to datacenter
  - ***Collaborate*** between multiple MNOs to fully release the power of data
- Federated learning can fulfill the above requirements
  - Temporal dependencies are modeled by local model, spatial dependencies are captured through model aggregation
  - Can be deployed at the edge sever
  - No need to transfer raw data, just model
  - Can be readily shared among different MNOs

Federated Learning

Central server

Round t

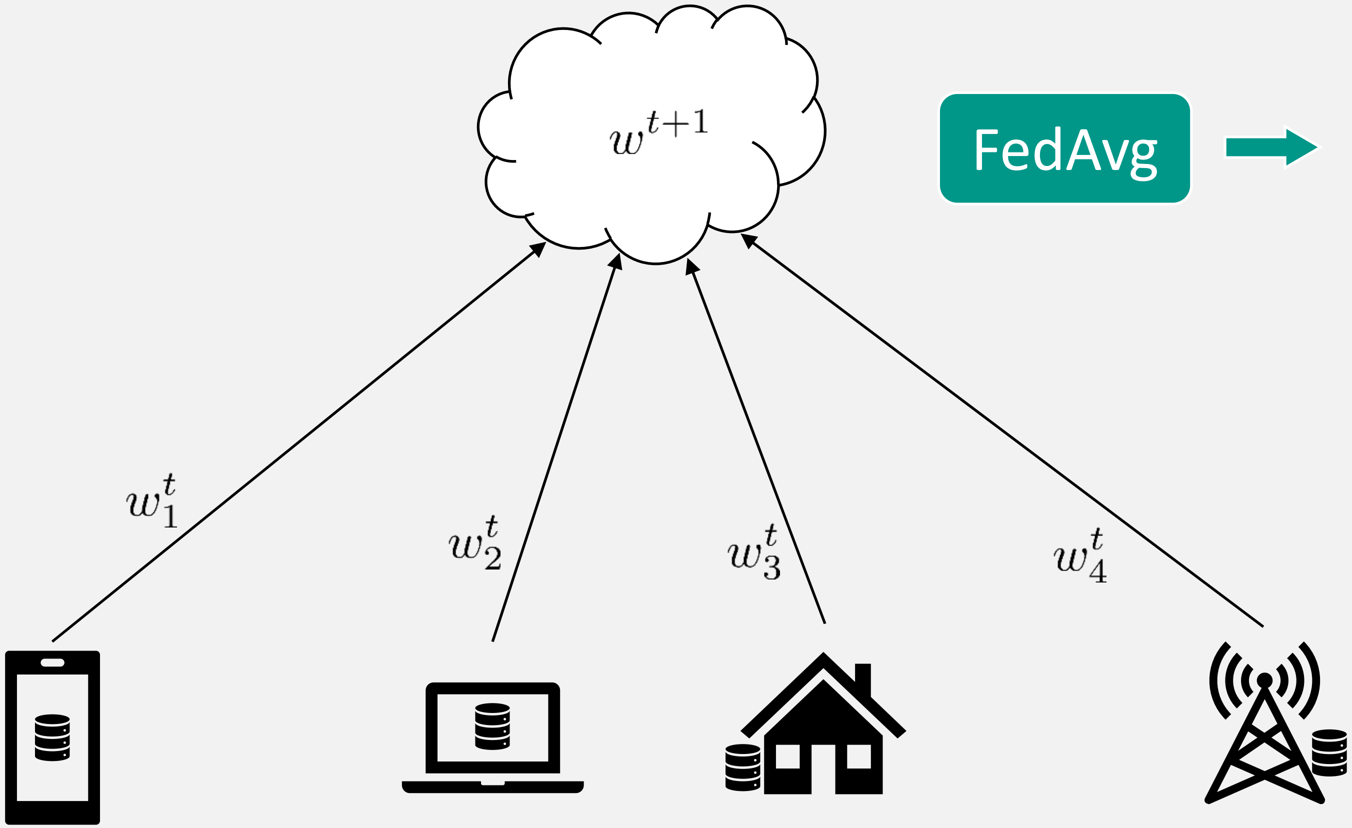
Local clients



Update w_t using local data and get new local model w_t^i

Federated Learning

Central server



FedAvg → $w^{t+1} = \sum_k w_k^t$

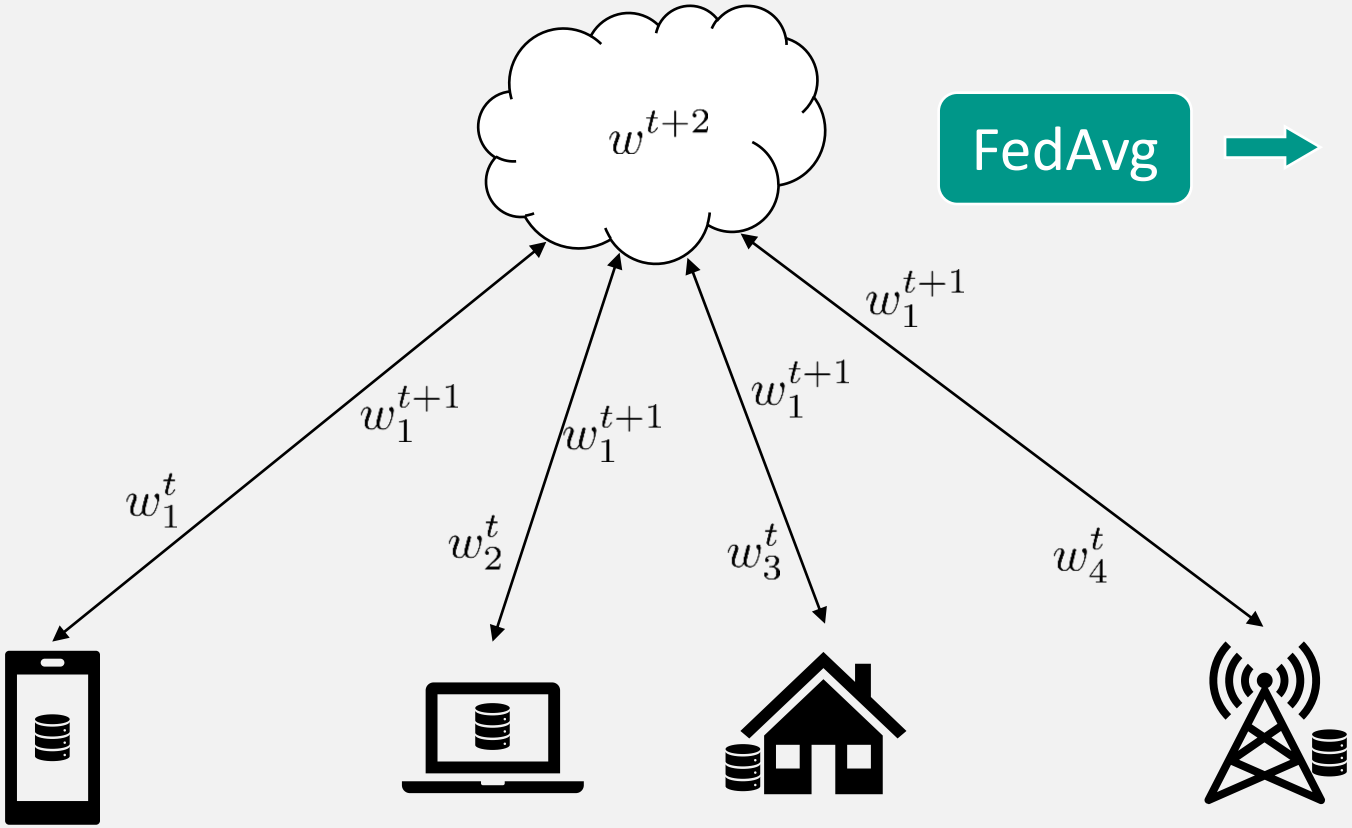
Round t

Local clients

Update w_t using local data and get new local model w_t^k

Federated Learning

Central server



FedAvg $\rightarrow$ $$w^{t+2} = \sum_k w_k^{t+1}$$

Round $t + 1$

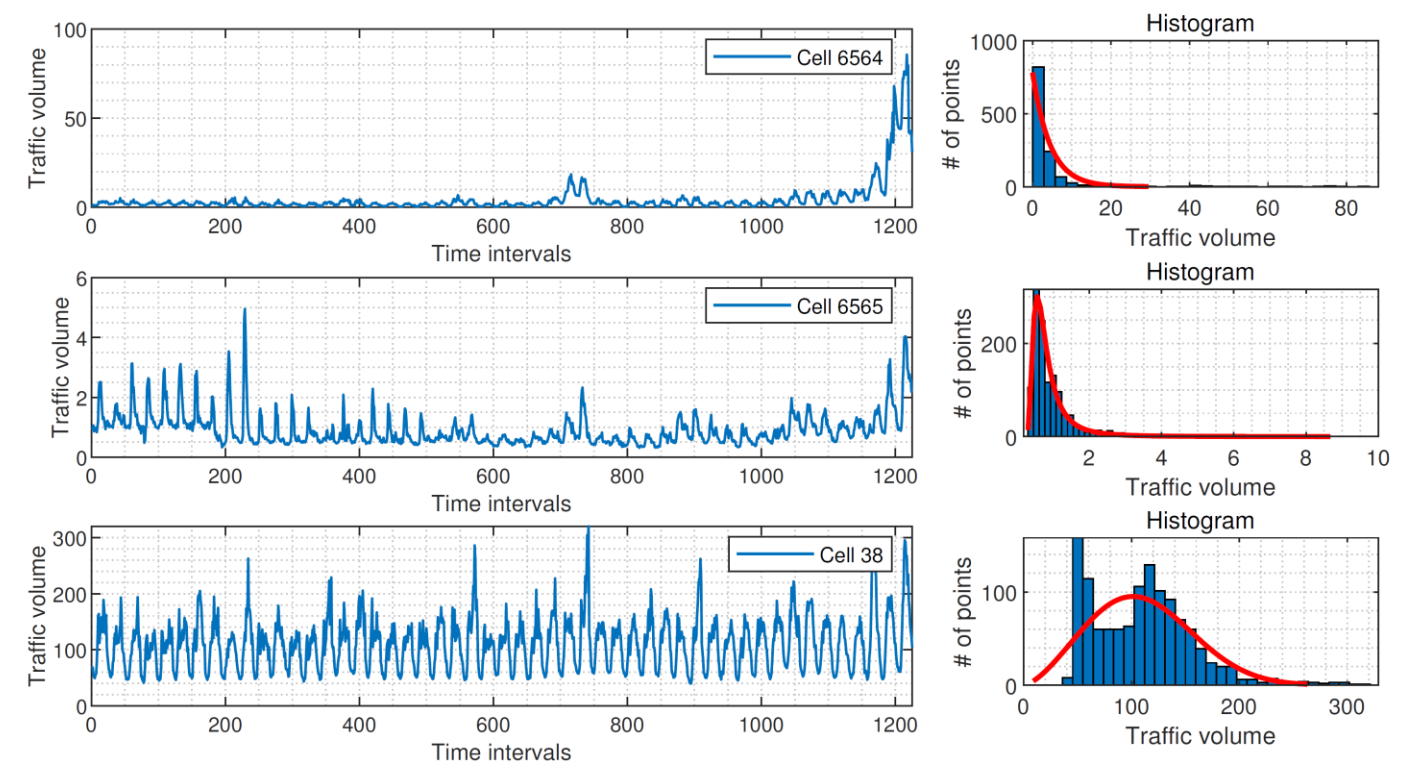
Local clients

Iterate the above process until end, then we get the final model

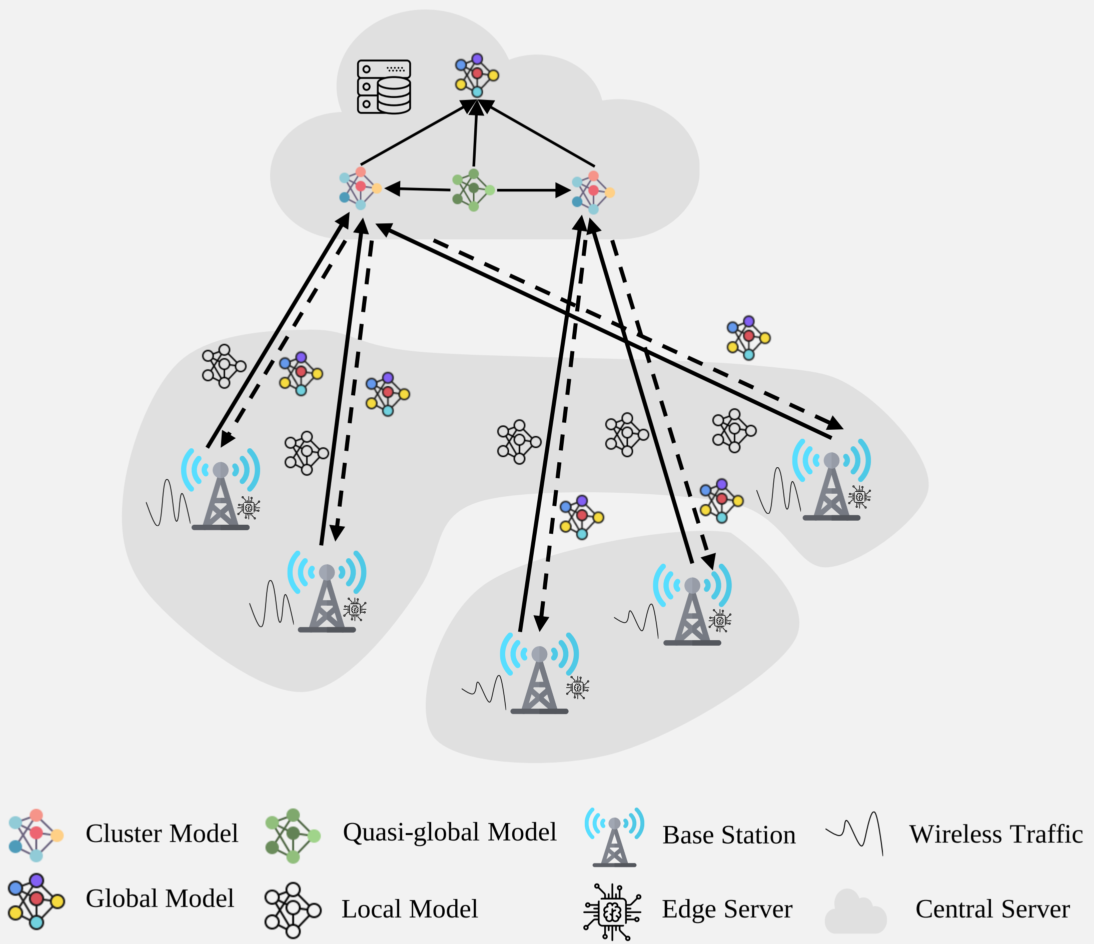
Update w_{t+1} using local data and get new local model w_{t+1}^k

FedAvg for Wireless Traffic Prediction

- It works, but suffers from precision problem
  - Wireless traffic data are ***highly heterogenous***, different places have different traffic patterns
  - ***Simple average*** of local model to produce the global one ***generalize not well***
- Motivation
  - Train a ***well-generalized global model*** by ***reducing heterogeneity*** of wireless traffic



System Model and Problem Formulation



\# of BSs

\# of clusters

\# of BSs in cluster c

Loss function

parameters

$$\min_{w\in\mathbb{R}^d} f(x) = \frac{1}{KC}\sum_{c=1}^{C}\sum_{m=1}^{K_m} F_{c,m}(w)$$

\# of data samples

Specific error function

$$F_{c,k}(w) = \sum_{i=1}^{N}\mathcal{L}(w; x_i, y_i)$$

$$y_i = x_t$$

$$x_i = \{x_{t-pL}, \cdots, x_{t-L}, x_{t-\tau}, \cdots, x_{t-1}\}$$

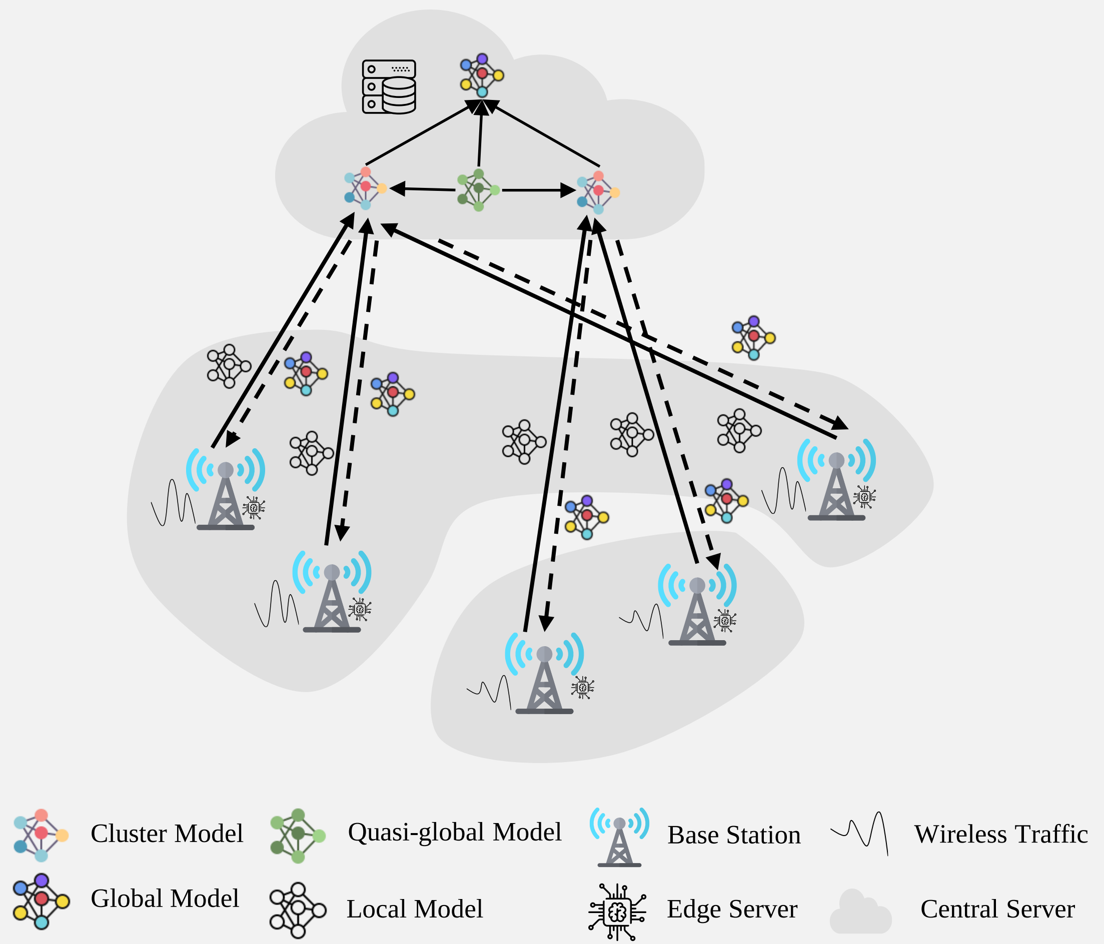
Length of periodicity

Period span

Length of closeness

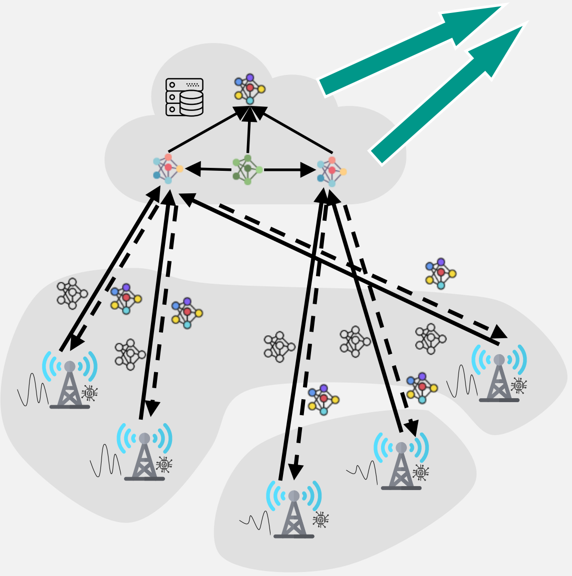
Time index

FedDA Workflow



Data augmentation	Each BS calculates its ***hourly statistical mean traffic value*** and send the data (φ%) to central
BS clustering	Central server train a quasi-global model and ***cluster the BSs into different groups***, based on the augmented
Model training	Perform global model training using ***dual attention-based optimization*** scheme

Dual Attention-Based Federated Optimization



$$\arg\min_{w^{t+1}} \left\{ \sum_{c=1}^{C} \frac{1}{2} \boxed{\alpha_c} \mathcal{L}(w^t, w_c^{t+1})^2 + \frac{1}{2} \rho \boxed{\beta} \mathcal{L}(w^t, w_Q)^2 \right\}$$

Local Attention

Layer-wise attention score computed via the distance between ***local*** model and the ***global*** model

Quasi-global Attention

Layer-wise attention score computed via the distance between ***quasi-global*** model and the ***global*** model

Intra-cluster update

Inter-cluster update

The global model has a minimum distance to each local model (***enhance personalization***) and quasi-global model (***reduce heterogeneity***) in parameter space.

Dual Attention-Based Federated Optimization

Local update $$w_c^{t+1} = w_c^t - \eta \nabla \mathcal{L}(w_c^t; x, y)$$

Sever update $$w^{t+1} = w^t - \gamma \{ \sum_{c=1}^{C} \alpha_c (w^t - w_c^{t+1}) + \rho \beta (w^t - w_Q) \}$$

Evaluation: Performance Comparisons

- Experiments on two real-world datasets

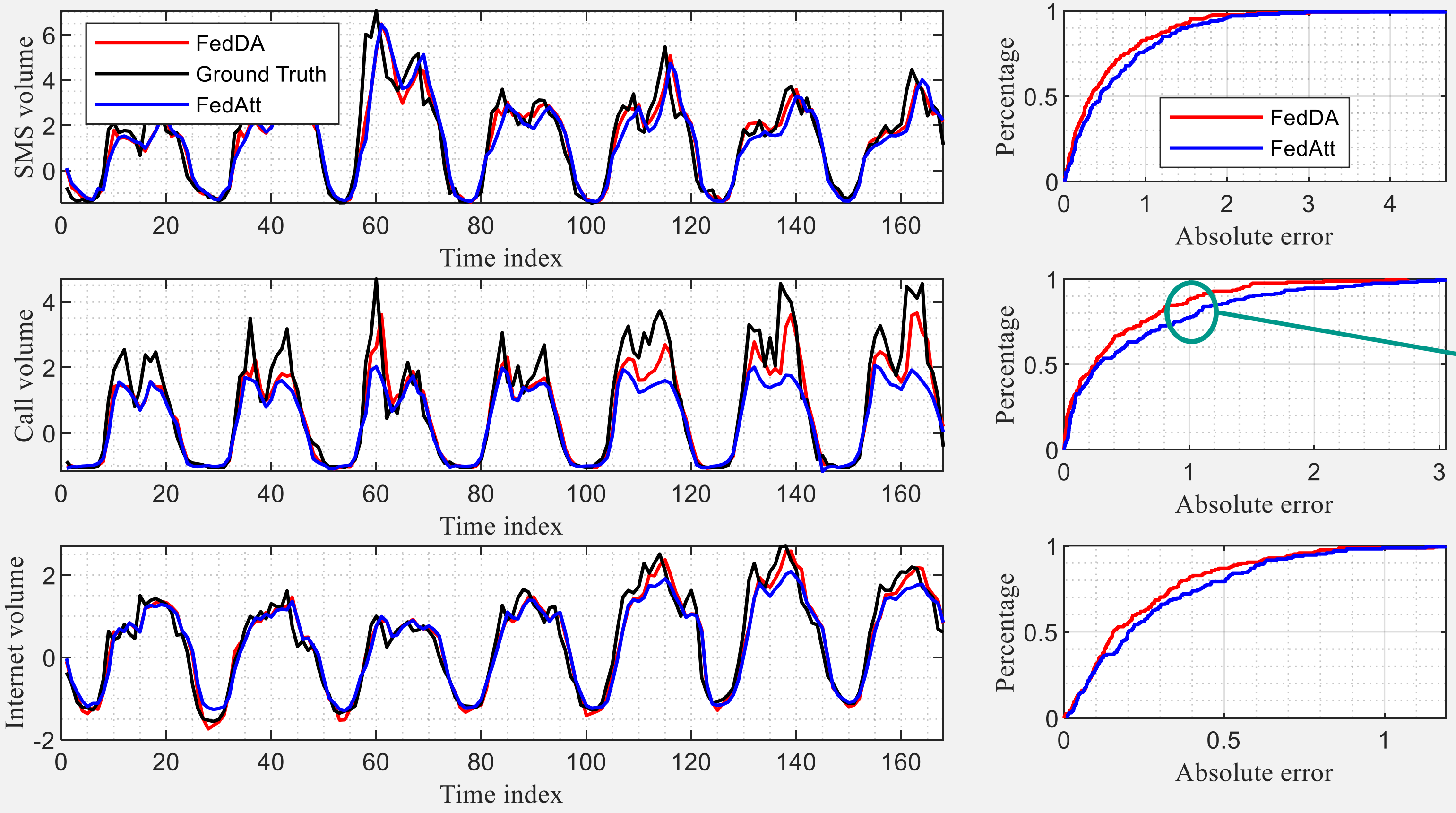
Methods	Milano						Trento					
	MSE			MAE			MSE			MAE		
	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet
Lasso	0.7580	0.3003	0.4380	0.6231	0.4684	0.5475	4.7363	1.6277	5.9121	1.3182	0.8258	1.5391
SVR	0.4144	0.0919	*0.1036*	0.3528	0.1852	*0.2220*	5.2285	1.7919	5.9080	1.0390	0.5656	1.0470
LSTM	0.5608	0.1379	0.1697	0.4287	0.2458	0.2936	3.6947	1.1378	4.6976	0.9426	0.5013	1.1193
FedAvg	0.3744	0.0776	0.1096	0.3386	0.1838	0.2319	2.2287	1.6048	4.7988	0.7416	0.5319	1.0668
FedAtt	0.3667	0.0774	0.1096	0.3375	*0.1837*	0.2321	2.1558	1.5967	4.7645	0.7444	0.5306	1.0629
FedDA (φ=1)	0.3559	*0.0752*	0.1118	0.3353	0.1820	0.2367	2.1468	1.4925	4.4335	0.7478	0.5140	1.0212
FedDA (φ=10)	*0.3481*	0.0753	0.1062	*0.3321*	0.1810	0.2275	*2.0719*	*1.1699*	*3.9266*	*0.7320*	*0.4543*	*0.9504*
FedDA (φ=100)	**0.3322**	**0.0659**	**0.1033**	**0.3214**	**0.1741**	**0.2211**	**1.9703**	**1.0592**	**2.4473**	**0.6920**	**0.4281**	**0.7471**
$\uparrow$ (φ=100)	+9.4%	+14.9%	+5.8%	+4.8%	+5.2%	+4.7%	+8.6%	+33.7%	+48.6%	+7.0%	+19.3%	+29.7%

Our method achieves the best prediction results

The more data shared, the better prediction performance

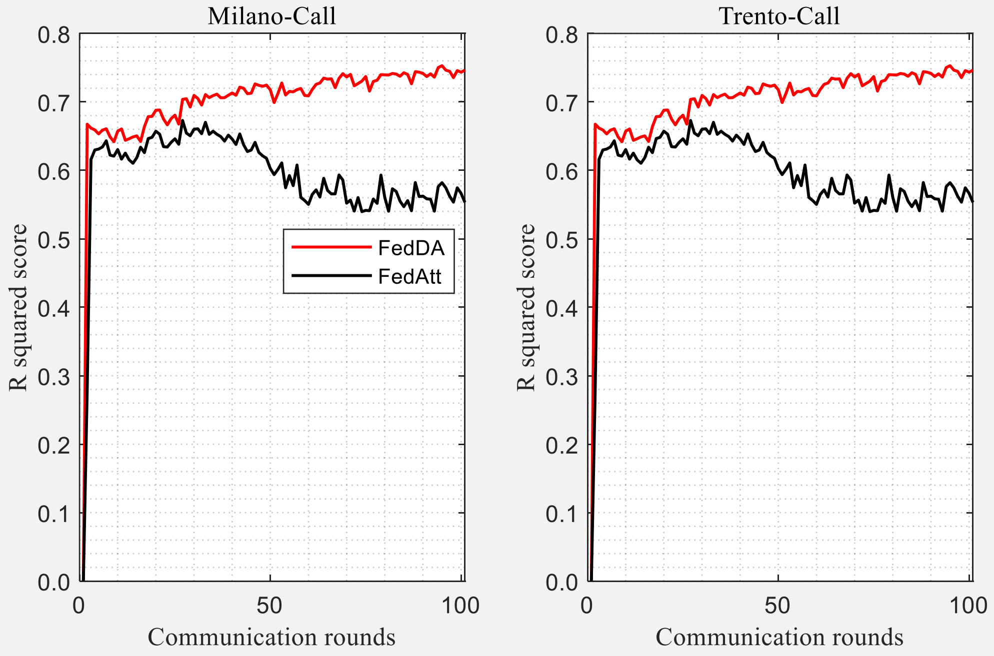
Evaluation: Predictions vs Ground Truth

FedDA achieved much better performance than baseline, especially when traffic values are large

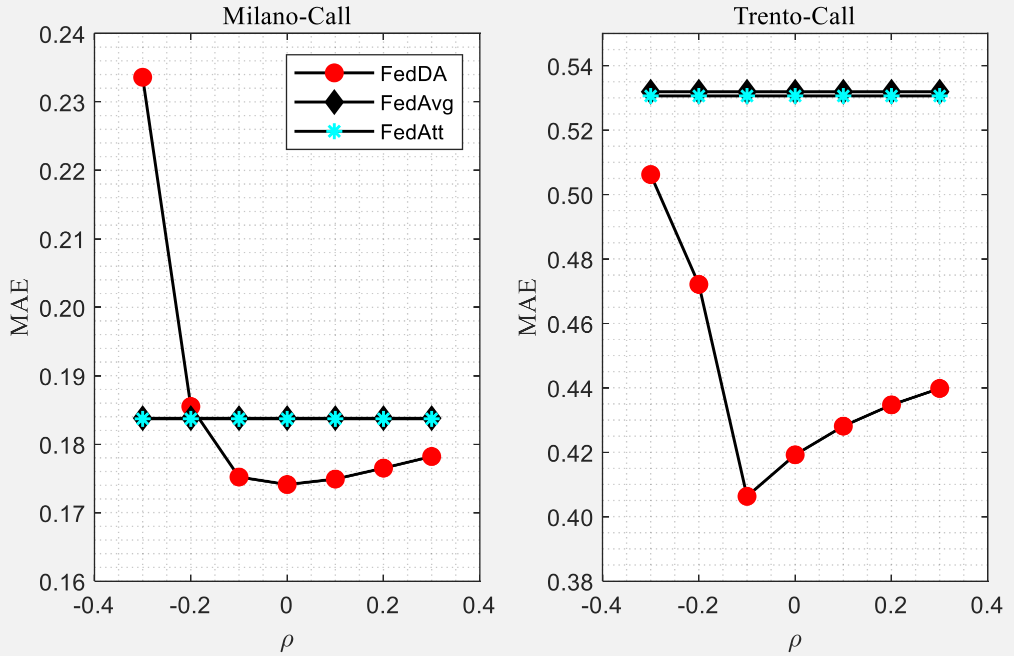


FedDA has a large portion of low errors

Evaluation: Accuracy vs Communication Rounds



FedDA can achieve higher prediction accuracy with fewer communications between local client and central server



Quasi-global attention (model) can indeed improve prediction performance

Summary

- We presented FedDA, a federated learning framework for wireless traffic prediction
- We designed an augmentation data sharing strategy to reduce data heterogeneity and a clustering strategy to enhance personalization
- We proposed a dual attention-based model aggregation scheme, which effectively balanced the global model’s personalization and generalization ability

IEEE International Conference on Computer Communications (INFOCOM 2021)

Thanks!

Code is available at https://github.com/chuanting/FedDA