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
Dual Attention-Based Federated Learning for Wireless Traffic Prediction

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

Round \( t \)

Update \( w_t \) using local data and get new local model \( w_t^i \)
Federated Learning

Round $t$

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

$w^{t+1} = \sum_k w_t^k$
Federated Learning

Central server

FedAvg

Round $t + 1$

Local clients

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

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

$$w^{t+2} = \sum_k w^{t+1}_k$$
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

Global Model
Cluster Model
Quasi-global Model
Local Model
Base Station
Edge Server
Wireless Traffic
Central Server

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

- \# of BSs
- \# of clusters
- \# of BSs in cluster \( c \)
- Loss function
- Parameters

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

- \# of data samples
- Specific error function

\[
x_i = \{x_{t-L}, \ldots, x_{t-L}, x_{t-\tau}, \ldots, x_{t-1}\}
\]

- Length of periodicity
- Period span
- Length of closeness
- Time index
FedDA Workflow

Each BS calculates its hourly statistical mean traffic value and send the data (φ%) to central

Central server train a quasi-global model and cluster the BSs into different groups, based on the augmented

Perform global model training using dual attention-based optimization scheme
Dual Attention-Based Federated Optimization

\[
\text{arg min}_{w^{t+1}} \left\{ \sum_{c=1}^{C} \frac{1}{2} \alpha_c \mathcal{L}(w^t, w_c^{t+1})^2 + \frac{1}{2} \rho \beta \mathcal{L}(w^t, w_Q)^2 \right\}
\]

**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.

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

$$w^{t+1} = w^{t} - \gamma \left\{ \sum_{c=1}^{C} \alpha_{c}(w^{t} - w_{c}^{t+1}) + \rho \beta (w^{t} - w_{Q}) \right\}$$
Evaluation: Performance Comparisons

- Experiments on two real-world datasets

<table>
<thead>
<tr>
<th>Methods</th>
<th>Milano</th>
<th>Trento</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td></td>
<td>SMS</td>
<td>Call</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.7580</td>
<td>0.3003</td>
</tr>
<tr>
<td>SVR</td>
<td>0.4144</td>
<td>0.0919</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.5608</td>
<td>0.1379</td>
</tr>
<tr>
<td>FedAvg</td>
<td>0.3744</td>
<td>0.0776</td>
</tr>
<tr>
<td>FedAtt</td>
<td>0.3667</td>
<td>0.0774</td>
</tr>
<tr>
<td>FedDA (γ=1)</td>
<td>0.3559</td>
<td>0.0752</td>
</tr>
<tr>
<td>FedDA (γ=10)</td>
<td>0.3481</td>
<td>0.0753</td>
</tr>
<tr>
<td>FedDA (γ=100)</td>
<td>0.3322</td>
<td>0.0659</td>
</tr>
</tbody>
</table>

|            | +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
FedDA achieved much better performance than baseline, especially when traffic values are large.
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
Thanks!

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