Efficient Wireless Traffic Prediction at the Edge: A Federated Meta-Learning Approach

Author 1, Author 2, and Author 3

***Abstract*—Wireless traffic prediction plays a vital role in managing high dynamic and low latency communication networks, especially in 6G wireless networks. Regarding data and computing resources constraints in edge devices, federated wireless traffic prediction has attracted considerable interest. However, federated learning is limited to dealing with heterogeneous scenarios and unbalanced data availability. Along this line, we propose an efficient federated meta-learning approach to learn a sensitive global model with knowledge collected from different regions. The global model can efficiently adapt to the heterogeneous local scenarios by processing only one or a few steps of fine-tuning on the local data sets. Additionally, distance-based weighted model aggregation is designed to capture the dependencies among different regions for better spatial-temporal prediction. We evaluate the performance of the proposed scheme by comparing it with the conventional federated learning approaches and other commonly used benchmarks for traffic prediction. The extensive simulation results reveal that the proposed scheme outperforms the benchmarks.**

***Index Terms*—Wireless traffic prediction, federated meta-learning, and heterogeneous scenarios**

I. Introduction

WITH the emergence of the concepts, such as 6G wireless networks [1], [2], Internet of Things (IoT), and unmanned aerial vehicle (UAV) assisted networks [3], the wireless traffic is anticipated to be dynamic, complex, and excessively high in scale. Wireless traffic prediction [4], [5] is one of the core ingredients for 6G networks since proactive resource allocation and green communications rely heavily on the accurate prediction of future traffic state. For example, adaptive channel assignment can be obtained by predicting wireless traffic to avoid traffic congestion [6].

Recently, deep learning (DL) approaches have achieved promising improvement for wireless traffic prediction [7], [8]. Recurrent neural network (RNN) is exploited in [9], [10] to perform spatial-temporal wireless traffic prediction. In [11], a spatial-temporal densely connected network (STDenseNet) is proposed for city-scale wireless traffic prediction. Zhang *et al.* exploit transfer learning to capture the complex patterns hidden in cellular data and transfer the knowledge to various traffic [4]. Transfer learning solves the problem of limited data and avoids training from scratch. However, the knowledge needs to be learnt and transferred from a region with a similar scenario.

The aforementioned centralized DL schemes [4], [11] need access to the geographically distributed data sets, which are hard to guarantee in wireless networks due to privacy concerns and communication overhead. Therefore, it naturally triggers the idea of federated learning (FL) solution for wireless traffic prediction, such as FedDA [12]. FL can significantly reduce the network bandwidth and latency by sending only the model parameters rather than the raw data stream. However, it is challenging to ensure good performance when FL applications face spatially-correlated scenarios [13]. FedDA adopted a clustering scheme to solve the spatial dependency modeling. But the clusters in FedDA are predetermined, which is too rigid to model the local spatial dependencies.

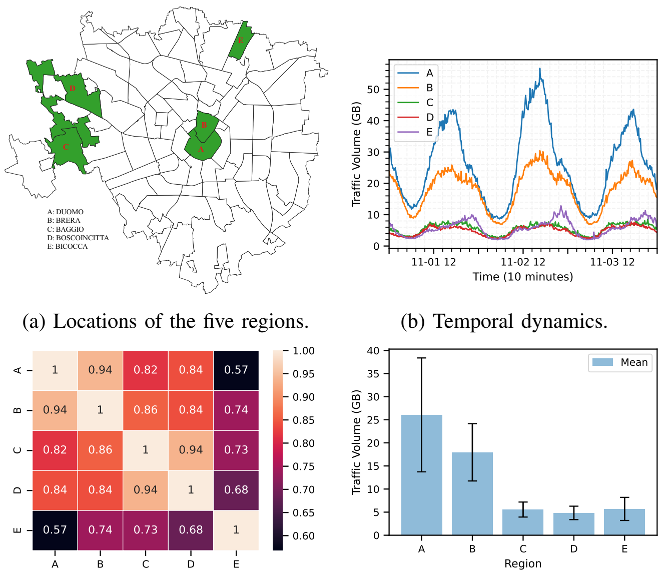
To avoid training from scratch and achieve personalized models for geographically distributed heterogeneous data set, data-sharing strategy [14], multi-task learning [10], and meta-learning [15] are adopted to overcome the statistical heterogeneity problem confronted with FL. But data-sharing breaks the principle of data privacy. Multi-task learning relies heavily on the assumption of certain task relationships, limiting its ability to solve the heterogeneity problem. Meta-learning model, on the other hand, is capable of well adapting or generalizing to new tasks and new environments. Thus, in this letter, we introduce model-agnostic meta-learning (MAML) into wireless traffic prediction under the FL framework to achieve efficient wireless traffic prediction at the edge. Specifically, we aim to train a sensitive initial model that can adapt fast to heterogeneous scenarios in different regions. A distance-based weighted model aggregation is further proposed and integrated to capture the dependencies among different regions for better spatial-temporal prediction. The proposed scheme inherits all the benefits from the FL architecture and guarantees the extra personalized characteristic to each local model.

II. Wireless Traffic Data and Problem Formulation

A. Wireless Traffic Data

The wireless traffic data sets are call detail records (CDRs) from the city of Milan, Italy and the province of Trentino, Italy, collected every 10 minutes over a two-month time span [16]. The raw CDRs are geo-referenced, anonymized and aggregated Internet traffic data based on the location of the regions. Specifically, a CDR record is logged if a user transfers more than 5 MB of data or spends more than 15 minutes online. After that, these records are grouped by administrative regions to protect privacy.

The patterns hidden in wireless traffic are complex and challenging to be modelled. The characteristics of wireless traffic in heterogeneous scenarios are analysed in Fig. 1, which includes the physical locations of five regions of Milan and the corresponding temporal and spatial traffic dynamics. We can observe that some regions have similar temporal patterns visually and high spatial correlation statistically. For example, region A and region B are physically near each other and have the same peak traffic hours. Their traffic series also have



(a) Locations of the five regions.

(b) Temporal dynamics.

(c) Spatial correlations.

(d) Traffic statistics.

Fig. 1: Spatial and temporal characteristics of wireless traffic.

high spatial correlations (0.94 in terms of Pearson correlation coefficient calculated with traffic vectors of A and B). But we also observe that some regions have distinct traffic patterns. For example, region A and region E have different peak traffic hours and small correlations. Besides, as shown in Fig. 1d, different regions have various traffic statistics. In this context, we need to train a model capable of capturing both the pattern similarity (spatial and temporal dependencies) and the pattern diversity (personalization).

B. Problem Formulation

We consider a decentralized communication network among geographically distributed regions. For each region, a local client records the wireless traffic and conducts the local model update. $\mathcal{C} = \{1, ..., k, ..., K\}$ denotes the clients set, where k is the index and K is the total number of the local clients. The sequential traffic data sets are divided into N time slots. In the n-th time slot, d_n is the random variable representing the traffic volume, and the closeness dependency $\mathbf{x}_n = \{d_{n-m}, d_{n-m+1}, ..., d_{n-1}\}$ is regarded as the input feature, where m is the number of the nearest data points taken into consideration. Suppose d_n to be the prediction target which can be labelled as the output y_n, since we consider the one-step-ahead prediction. Thus the input-output pair $\{\mathbf{x}_n, y_n\}$ can be obtained by using sliding window scheme.

The samples are locally generated. The number of samples N^k varies from client to client, and zero sample is possible for an individual client. Furthermore, the training set of the k-th client $\mathcal{P}_k$ is divided into support set $\mathcal{P}_k^s$ and query set $\mathcal{P}_k^q$. Personalized knowledge is preserved and internally transferred via $\mathcal{P}_k^s$ to $\mathcal{P}_k^q$. To aggregate the local models at the central server and inherit the globe model from the central server to each local client, the uplinks and the downlinks between the local clients and the central server are built up.

Generally, the objective of FL-based traffic prediction is to obtain a global model with parameter θ that can minimize the average loss function of the local data sets, which is denoted as

$$\min_{\theta} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(\theta; \mathcal{P}_k), \tag{1}$$

where $\mathcal{L}(\theta; \mathcal{P}_k)$ is the loss function representing the differences between the predicted traffic volume $\hat{y}_n$ and the ground truth y_n. Taking the mean squared error (MSE) as the metric for example, the loss function is defined as

$$\mathcal{L}(\theta; \mathcal{P}_k) = \frac{1}{N_k} \sum_{\{\mathbf{x}_n, y_n\} \in \mathcal{P}_k} (\hat{y}_n - y_n)^2. \tag{2}$$

In contrast to the traditional FL-based traffic prediction targeting to train an ordinary model that ingests all clients, our objective is to obtain a sensitive global model that can adapt fast to a heterogeneous distribution of scenarios. In this regard, we manage to minimize the loss between the true value of each client's traffic and the predicted value calculated based on the fine-tuned model, which is obtained by proceeding one or a few steps of fine-tuning on part of the local data set. The objective is formally described as

$$\min_{\theta} \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}(\theta - \alpha \sum_{j=0}^{J-1} \nabla_{\theta} \mathcal{L}(\theta_{k,j}; \mathcal{P}_k^s); \mathcal{P}_k^q), \tag{3}$$

where $\nabla_{\theta} \mathcal{L}(\theta_{k,j}; \mathcal{P}_k^s)$ denotes the gradient corresponding to the j-th steps of local update, $j \in [0, J)$. $\theta - \alpha \sum_{j=0}^{J-1} \nabla_{\theta} \mathcal{L}(\theta_{k,j}; \mathcal{P}_k^s)$ is the model obtained after fine-tuning on support sets $\mathcal{P}_k^s$. The intuition of problem (3) is to minimize the average loss of the fine-tuned models proceeding on the query sets $\mathcal{P}_k^q$, which would be the new tasks we expect our model to fast adapt to.

III. Federated meta-learning Approach

In this section, we propose a federated meta-learning approach for wireless traffic prediction. The training system is configured with a decentralized structure, the same as the conventional FL-based approach [12]. We implement the MAML strategy in the federated framework and conduct distance-based weighted model aggregation to simultaneously achieve efficient and personalized traffic prediction. Once the global model is well trained, the test is conducted individually at the edge after a few steps of gradient descent fine-tuning. The scheme is illustrated in Algorithm 1.

A. MAML-Enhanced Parameter Learning

We randomly initialize the global model parameter θ. A set of $C = \max(\delta K, 1)$ clients denoted as $\mathcal{C}_t$ is randomly selected during each training episode, where δ is the hyper-parameter qualifying the fraction of the clients chosen at each round. For each client $c \in C_t$, we load the current global model in parallel and initialize the local model parameter $\theta_{c,0}^t$ by reproducing the global model parameter θ^t. Thereafter, a batch of traffic prediction tasks $\mathcal{T}_c^s$ is sampled from the support set $\mathcal{P}_c^s$. J steps of gradient descent are conducted on sampled $\mathcal{T}_c^s$, and the updated model is internally transferred to preserve the personalized knowledge. Formally, the local model parameters updated at the j-th step are calculated as follows

$$\theta_{c,j}^t = \theta_{c,j-1}^t - \alpha \nabla_{\theta^t} \mathcal{L}(\theta_{c,j-1}^t; \mathcal{T}_c^s). \tag{4}$$

Algorithm 1: Federated Meta-learning Algorithm for Wireless Traffic Prediction

```
   Input: data sets $\mathcal{P}$, step size parameters $\alpha$ and $\beta$,
          fraction of selected client $\delta$
   Output: Learned model parameters $\theta$
1  random initialize $\theta$
2  for each round $t = 0, 1, 2, \cdots$, do
3      $C = \max(\delta K, 1)$
4      Sample a set $\mathcal{C}_t$ of $C$ clients
5      for each client $c \in \mathcal{C}_t$ in parallel do
6          Load global model: $\theta^t_{c,0} = \theta^t$
7          Sample a batch of tasks $\mathcal{T}^s_c$ from $\mathcal{P}^s_c$
8          for each step $j = 1, 2, \cdots, J$ do
9              $\theta^t_{c,j} = \theta^t_{c,j-1} - \alpha \nabla_{\theta^t} \mathcal{L}(\theta^t_{c,j-1}; \mathcal{T}^s_c)$
10         Sample a batch of tasks $\mathcal{T}^q_c$ from $\mathcal{P}^q_c$
11         Update model with (5)
12     Individual model enhancement based on spatial
         dependencies $\tilde{\theta}^{t+1}_c = \sum_{r \in \mathcal{C}_t} \tilde{\rho}^{t+1}_{c,r} \theta^{t+1}_{r,0}$
13     Global model update: $\theta^{t+1} = \frac{1}{C} \sum_{c \in \mathcal{C}_t} \tilde{\theta}^{t+1}_c$
```

Subsequently, a batch of tasks $\mathcal{T}^q_c$ is sampled from the query set $\mathcal{P}^q_c$. The local model is improved by rapidly adjusting to the sampled query tasks. The updated local models are uploaded to the central server for the knowledge integration from the heterogeneous scenarios, such as

$$\theta^{t+1}_{c,0} = \theta^t_{c,0} - \beta \nabla_{\theta^t} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c), \tag{5}$$

where $\nabla_{\theta^t} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c)$ is the second-order gradient descent conducted on the query tasks and is merged into the current local models corresponding to the slightly updated and internally transferred model parameter $\theta^t_{c,J}$. Taken equation (4) into consideration, the second-order gradient descent operation is given as

$$\begin{aligned}
\nabla_{\theta^t} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c) =& \nabla_{\theta^t_{c,J}} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c) \cdot \nabla_{\theta^t} \theta^t_{c,J} \\
=& \nabla_{\theta^t_{c,J}} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c) \cdot \nabla_{\theta^t_{c,J-1}} \theta^t_{c,J} \cdot \nabla_{\theta^t} \theta^t_{c,J-1} \\
=& \nabla_{\theta^t_{c,J}} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c) \cdot \prod_{j=1}^{J} \nabla_{\theta^t_{c,j-1}} \theta^t_{c,j} \\
=& \nabla_{\theta^t_{c,J}} \mathcal{L}(\theta^t_{c,J}; \mathcal{T}^q_c) \\
& \cdot \prod_{j=1}^{J} (\mathcal{I} - \alpha \nabla_{\theta^t_{c,j-1}} \nabla_{\theta^t} \mathcal{L}(\theta^t_{c,j-1}; \mathcal{T}^q_c)).
\end{aligned} \tag{6}$$

B. Distance-Based Weighted Model Aggregation

To further model the spatial dependencies among different regions, we propose a distance-based weighted model aggregation scheme. More specifically, once the central server received all the gradient information from the chosen clients at the t-th communication round, we calculate the cosine similarities among different regions, which yields a distance matrix $\boldsymbol{\rho}^{t+1}$

$$\boldsymbol{\rho}^{t+1} = \begin{bmatrix} \rho^{t+1}_{1,1} & \rho^{t+1}_{1,2} & \cdots & \rho^{t+1}_{1,C} \\ \rho^{t+1}_{2,1} & \rho^{t+1}_{2,2} & \cdots & \rho^{t+1}_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t+1}_{C,1} & \rho^{t+1}_{C,2} & \cdots & \rho^{t+1}_{C,C} \end{bmatrix}, \tag{7}$$

where $\rho^{t+1}_{c,r}$ measures the cosine similarity between region c and region r, and is computed as

$$\rho^{t+1}_{c,r} = \frac{\theta^{t+1}_{c,0} \cdot \theta^{t+1}_{r,0}}{||\theta^{t+1}_{c,0}|| \cdot ||\theta^{t+1}_{r,0}||}. \tag{8}$$

For each client c, an enhanced individual model incorporating spatial dependencies is obtained as

$$\tilde{\theta}^{t+1}_c = \sum_{r \in \mathcal{C}_t} \tilde{\rho}^{t+1}_{c,r} \theta^{t+1}_{r,0}, \tag{9}$$

where $\tilde{\rho}^{t+1}_{c,r}$ is the softmax version of $\rho^{t+1}_{c,r}$. Then, the central server update the global model as follows

$$\theta^{t+1} = \frac{1}{C} \sum_{c \in \mathcal{C}_t} \tilde{\theta}^{t+1}_c. \tag{10}$$

The above induced global model captures the spatial dependencies among different regions and can adapt to new traffic patterns. Notice that, for fairness consideration, we set the size of $\mathcal{T}_c$ sampled in every client identical.

C. Model Personalization with Local Adaption

Before adopting the model to a new traffic prediction task of a specific client, fine-tuning is executed on the local data set to adjust the model to the private data of each client. Specifically, we sample a batch of tasks $\mathcal{T}^s_c$ from the local data set and conduct only one or a few gradient descent steps. The above mentioned adaption is the repetition of the sampling and internal updating process (line 7-9) in Algorithm 1. The volume of the traffic is predicted by implementing the personalized mode with parameter $\theta_{c,J}$, which is expressed as

$$\theta_{c,J} = \theta - \alpha \sum_{j=0}^{J-1} \nabla_\theta \mathcal{L}(\theta_{c,j}; \mathcal{T}^s_c). \tag{11}$$

The model can be evaluated in terms of MSE based on test data sets $\mathcal{P}^{\text{test}}_c$, such as

$$\mathcal{L}(\theta_{c,J}; \mathcal{P}^{\text{test}}_c) = \frac{1}{N^{\text{test}}_c} \sum_{\{\mathbf{x}_n, y_n\} \in \mathcal{P}^{\text{test}}_c} (\hat{y}_n - y_n)^2, \tag{12}$$

where N^{test}_c is the number of samples for testing.

IV. Experimental Results and Analysis

This section gives a detailed introduction of the experimental settings, baseline methods, and evaluation metrics. After that, we analyze and report the achieved experimental results.

A. Experiment Settings

Our experiment uses the first seven weeks' data to train a model and the last week's data to test the model. In each communication round we assume only a few clients, e.g., δK, are involved, and we set δ to 0.1. K equals to 88 and 223 for the Milano and Trentino, respectively. To generate data samples, the window size m is set to 6. Data samples are standardized to accelerate the training speed. We design a neural network with L layers, and each layer has M neurons. Considering the amount of data in each region and the power restrictions of the edge server, L and M are set to 3 and 40, unless otherwise specified. We train our model for 100 consecutive rounds with batch size 20 by using SGD. The choices of learning rates, i.e., α and β, are obtained by a grid search over $\alpha, \beta \in \{0.1, 0.01, 0.001\}$.

TABLE I: Prediction comparisons among different algorithms.

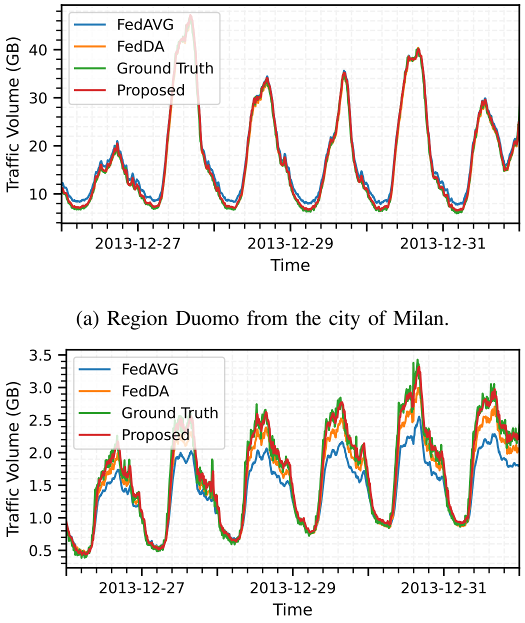
	Milano		Trentino	
Methods	MSE	MAE	MSE	MAE
HA	1.2839	0.9939	12.0363	2.3288
SVR	0.0187	0.0883	10.0037	1.5755
RF	0.0218	0.0918	3.5385	0.9296
FedAvg	0.0196	0.0965	1.1033	0.5834
FedDA	0.0179	0.0816	0.5463	0.3933
Proposed-s	0.0170	0.0803	**0.4815**	**0.3544**
Proposed-w	0.0170	0.0790	0.5498	0.3812
Proposed-d	**0.0169**	**0.0782**	0.5258	0.3855

B. Baselines and Evaluation Metrics

We compare our algorithm with historical average (HA), support vector regression (SVR), random forest (RF), federated averaging (FedAvg) and FedDA. The first one is a classical time series prediction method. SVR and RF are traditional machine learning methods for wireless traffic prediction. FedAvg trains a global model by averaging the local ones. FedDA captures the spatial dependencies of regions by clustering. All baselines are trained in a fully distributed way except FedAvg and FedDA, which are trained in a federated manner. To make a fair comparison, FedAvg, FedDA, and our proposed method share exactly the same network architecture and are configured with the same (hyper-)parameters, e.g., learning rate and batch size. To explore our model's robustness, we consider three variants. i.e., a standard network with L=3 and M=40 named Proposed-s, a wide network with L=3 and M=400 named Proposed-w, and a deep network with L=10 and M=40 named Proposed-d. We evaluate the prediction performance of different algorithms in terms of MSE and mean absolute error (MAE) metrics.

C. Prediction Results

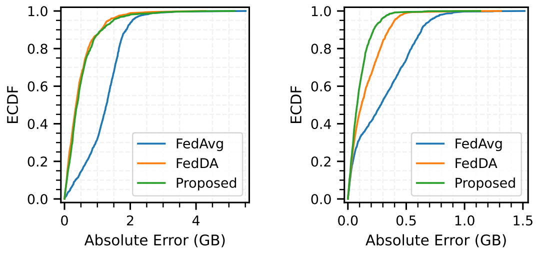
We repeat the experiments 10 times and report the averaged quantitative prediction results in Table I. The best results are marked in bold for clearness. We can observe from this table that our proposed method achieves the best prediction results, which validates the effectiveness of MAML integrated into a federated framework. Take the MSE as an example, among all the baselines, the best results are 0.0179 and 0.5463, on the Milano and Trentino data set, respectively. Our method's best results are 0.0169 and 0.4815. Thus, we can see a clear performance improvement, especially for the Trentino data set. The improvement percentages are up to 5% and 11.9% for Milano and Trentino data sets, respectively. In addition, we notice that our method is relatively robust to network architecture as all these three variants achieve similar prediction results. The performance of HA is generally poor as they are parameter-free and have no ability to learn the hidden patterns. Learning-based fully distributed methods can usually achieve lower prediction errors than HA, as they can model the traffic dynamics through adjustable parameters. Besides, a model's prediction ability has a positive relationship with the number of parameters. Another thing worth noting is that FL-based methods are superior to fully distributed methods since they involve model aggregation and can fuse knowledge of different regions. This is an effective way to capture the spatial dependencies among different regions. But our proposed method achieves better predictions than FedAvg and FedDA, since it is aware of the spatial dependency diversities among different regions.



(a) Region Duomo from the city of Milan.

(b) Region Andalo from Trentino province.

Fig. 2: Predictions versus ground truth values.



(a) Duomo

(b) Andalo

Fig. 3: ECDF as a function of absolute error.

D. Prediction vs Ground Truth

We report region-level prediction results in this subsection. For each data set, a random region is selected and the comparisons between predictions and ground truth values are plotted in Fig. 2. Besides, the empirical cumulative distribution function (ECDF) of absolute prediction errors is also reported, and the results are summarized in Fig. 3. We include the results of FedAvg and FedDA in both figures for comparisons.

From Fig. 2, we observe that our method has similar performance with FedAvg and FedDA on the Milano data set. But on the Trentino data set, our method achieves much better results than FedAvg and FedDA, especially when the traffic volume increases from the fourth day. Our method's superiority can be more clearly reflected by the ECDF of prediction errors in Fig. 3. For example, on the Trentino data set, 82.5% prediction errors of our method are less than 0.2 GB. While the cases

TABLE II: MSE comparisons of different models when dealing with different traffic scenarios.

Scenarios	Target Region	Methods		
		FedAVG	FedDA	Proposed
Homogeneous	Ala	0.2295	0.2372	0.2240
	Avio	0.1230	0.1264	0.1215
	Trambileno	0.0995	0.1097	0.0979
	Bosentino	0.6514	0.6629	0.6297
Heterogeneous	Pellizzano	2.4173	2.2625	**1.6113**

for FedAVG and FedDA are 40% and 65%, respectively. The results indicate that introducing MAML and distance-based weighted model aggregation into federated learning can indeed enhance the generalization ability of the global model, particularly for high heterogeneous scenarios, such as the Trentino data set.

E. Homogeneous vs Heterogeneous

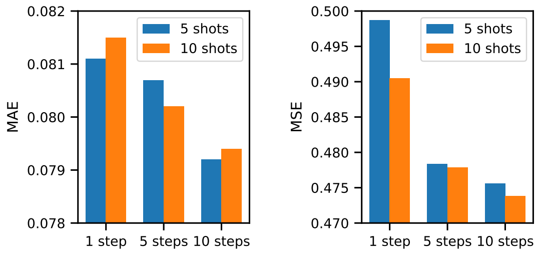
To further demonstrate the ability of different models when dealing with heterogeneous wireless traffic, we select four regions with high similarities and another one with distinct traffic patterns with the other four. We train a model using data samples from the former four regions and test the model's performance on these five regions separately to mimic homogeneous and heterogeneous scenarios. The selected regions and obtained results of FedAVG, FedDA, and our proposed method are summarized in Table II. We can clearly notice from Table II that all three methods perform fairly well for the homogeneous scenarios. But when dealing with heterogeneous data set, i.e., test a model on data samples of unseen (possibly unbalanced) and distinct traffic patterns, our proposed method achieves much better performance compared with FedAVG and FedDA. Thus, the results in Table II demonstrate the excellent adaptive ability of our method of dealing with heterogeneous wireless traffic datasets.

F. Impacts of hyper-parameters

There are two key hyper-parameters in our method, i.e., the number of data samples per slot and the fine-tuning steps. We report the results when varying these two hyper-parameters in Fig. 4. It can be seen from Fig. 4 that when the number of adaption steps or the number of data samples per slot increases, the performances of our method are improved since more data samples are involved in the local model update. But when the number of fine-tuning steps is large enough, the performance gain is minimal. In reality, the optimal choices of these two parameters can be obtained through a grid search scheme.

V. Conclusion

In this paper, we proposed efficient federated meta-learning approach for the decentralized wireless traffic prediction. Distance-based weighted model aggregation scheme was integrated to capture the spatial-temporal characterizes. By implementing the approach, we obtained a sensitive global model that can quickly adapt to heterogeneous scenarios and unbalanced data availability at the edge clients via only a few steps of fine-tuning. Three measures on two different data sets evaluated the effectiveness and efficiency of the approach. The impacts of hyper-parameters were also reported.



(a) Milano

(b) Trentino

Fig. 4: Parameter sensitivity.

The experimental results showed that our proposed approach outperforms other federated learning approaches and classical prediction methods.

References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, Oct. 2019.

[2] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[3] L. Zhang, A. Celik, S. Dang, and B. Shihada, "Energy-efficient trajectory optimization for uav-assisted IoT networks," *IEEE Transactions on Mobile Computing*, Apr. 2021.

[4] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, Mar. 2019.

[5] Y. Xu, F. Yin, W. Xu, J. Lin, and S. Cui, "Wireless traffic prediction with scalable gaussian process: Framework, algorithms, and verification," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1291–1306, Mar. 2019.

[6] F. Tang, B. Mao, Z. M. Fadlullah, and N. Kato, "On a novel deep-learning-based intelligent partially overlapping channel assignment in SDN-IoT," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 80–86, 2018.

[7] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2224–2287, Mar. 2019.

[8] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," in *Proc. ACM Mobihoc*, Los Angeles, CA, USA, Jun. 2018, pp. 231–240.

[9] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *Proc. IEEE INFOCOM* , Atlanta, GA, USA, May. 2017, pp. 1–9.

[10] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, "Spatio-temporal wireless traffic prediction with recurrent neural network," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 554–557, Jan. 2018.

[11] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656–1659, May 2018.

[12] C. Zhang, S. Dang, B. Shihada, and M.-S. Alouini, "Dual attention based federated learning for wireless traffic prediction," in *Proc. IEEE INFOCOM*, Vancouver, BC, Canada, May 2021, pp. 1–10.

[13] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 41–47, Dec. 2020.

[14] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, Jun. 2018.

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. PMLR ICML*, Sydney, Australia, Aug. 2017, pp. 1126–1135.

[16] G. Barlacchi, M. D. Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Scientific Data*, vol. 2, no. 1, pp. 1–15, Oct. 2015.