2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00032

A General Spatiotemporal Imputation Framework for Missing Sensor Data

Aabila Tharzeen[1], Sai Munikoti[2], Punit Prakash[1], Jungkwun Kim[3], and Balasubramaniam Natarajan[1]

[1]Kansas State University, Kansas, USA, [2]Pacific Northwest National Lab
[3]University of North Texas, TX, USA

***Abstract*—Many applications from precision agriculture, environmental monitoring and transportation networks rely on data collected across space and time over a large geographic area. Missing data can be a significant issue in these spatiotemporal databases, as it can reduce the accuracy of downstream data analysis, inferencing and control algorithms. Data imputation or the estimation of missing data can help fill these gaps by utilizing inherent spatial relationships and temporal patterns. However, existing approaches for estimating this missing information do not effectively capture all dimensions of the spatiotemporal data structure, resulting in erroneous predictions and poor performance. In this paper, we introduce a general framework that leverages a spatiotemporal graph constructed from the sensor network graph and temporal sensor data to capture the joint space-time dependencies. Specifically, we propose a graph neural network-based model in conjunction with a recurrent neural network to impute missing information and demonstrate the effectiveness of our approach for downstream tasks. Experiments on a traffic sensor network reveal enhanced imputation accuracy and up to 69% reduction in mean absolute error and 61% reduction in root mean square error compared to state-of-the-art imputation frameworks.**

***Index Terms*—Spatiotemporal imputation, GNN, LSTM, Traffic data**

I. Introduction

The emergence of IoT-based systems empowered by advances in sensing, communication and control has triggered a wealth of applications that rely on data collected across space and time. Examples of applications that rely on spatiotemporal data from sensor networks include power systems, traffic networks, air quality monitoring, and precision agriculture. While the integrated spatial and temporal information can lead to more efficient data analysis, the spatiotemporal data often contain missing observations due to various factors such as malfunctioning sensors or communication errors [1]. The presence of missing data can significantly reduce the accuracy of downstream tasks such as classification, clustering, and forecasting, leading to unreasonable inferences. Therefore there is a need to develop effective missing data imputation strategies that can be used in the preprocessing step or develop models that are robust to missing data.

A. Related work

A variety of spatiotemporal imputation models have been developed to address missing data in spatiotemporal datasets. Some of the classic statistical methods involve interpolation-based methods that use linear interpolation to estimate missing values based on the values from the neighboring time/spatial points [2], [3].

However, these classical methods rely on the assumption that the underlying data follows a smooth trend and fail to provide accurate estimates when there is a large number of missing points in the data. Yet another method to impute the missing data is by estimating/assuming the correlation between multiple variables in the dataset [4]. This approach requires a high degree of statistical expertise and can be computationally expensive [5].

Multilinear tensor completion that uses a low-rank tensor approximation based on observed entries to reconstruct the missing values is proposed in [6]. While the approach in [6] effectively captures multi-dimensional structural dependencies, it is unsuitable for complex interactions and diverse data missing patterns. A convolutional neural network based tensor completion (CoSTCo) method was proposed in [7]. While CoSTCo captures the non-linear relationships in the dataset, the transductive nature makes the algorithm less scalable. Imputation techniques based on machine learning algorithms use k-nearest neighbors [8] or support vector machines to estimate missing values based on patterns in the data [9]. These methods do not attempt to capture the complex relationships inherent in the spatiotemporal data and only rely on data similarity metrics.

Recently, deep learning-based approaches have been proposed to impute missing data. Denoising stacked autoencoder (DSAE) [10], is a typical deep learning model that combines denoising and autoencoders for imputation. However, DSAE does not account for the underlying spatial correlations. To leverage the spatial correlations, [11] proposes a multi-range convolutional neural network (CNN) to model spatial correlations and impute missing information. Though the method proposed in [11] effectively handles correlations in Euclidean space, they are inefficient in modeling relationships in non-Euclidean spaces. Recently, graph structures for relational reasoning have been utilized in Graph Convolutional Networks (GCN) [12], [1]. Though GCN is effective in modeling topological relationships, it is not tailored to capture temporal dependencies. Alternatively, recurrent neural network-based models can impute time series with missing values. Specifically, long short-term memory (LSTM) networks can capture and maintain long-term dependencies [13].

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00032

B. Contributions

In this work, we propose a novel inductive framework (G-LSTM) for missing data imputation that integrates a graph neural network with LSTMs to effectively capture both spatial and temporal dependencies. We use GraphSAGE [14] as a GNN module which is an inductive method and computationally efficient compared to GCN. Furthermore, GraphSAGE allows for incorporating node features in the embedding generation process, which can be useful for tasks where node attributes are essential. The proposed general framework can be used for both data imputation and prediction.

The performance of the proposed framework is highlighted using comprehensive case studies on real-world traffic datasets. The case studies include missing rates ranging from 10 % to 90%. Experimental results demonstrate that the proposed GNN integrated with the LSTM framework achieves improved imputation and maintains steady performance even when there are extreme missing conditions in comparison with the state-of-the-art imputation framework CoSTCo. The simulation results on the traffic network show up to 69% reduction in mean absolute error and 61% reduction in root mean square error when compared to state-of-the-art imputation framework.

The remainder of this paper is structured as follows. Section III introduces the imputation problem description and the required background. The proposed approach is explained in section IV and the results are discussed in section V. Finally, section VI concludes the paper.

II. Problem Statement

We represent the spatiotemporal dataset as an undirected graph $\mathbf{G} = (\mathbf{A}, \mathbf{E}, \mathbf{X}_t)$, where $\mathbf{A} \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ denotes the spatial adjacency matrix; N denotes the number of sensors and the element $\mathbf{a_{i,j}} \in \mathbf{A}$ is 1 if the sensors i and j are adjacent; Edge $\mathbf{E}$ represents the spatial relationship between the sensors, and $\mathbf{X}_t \in \mathbb{R}^{\mathbb{N}\times\mathbb{D}}$ is the set of $\mathbf{D}$ dimensional dynamic features measured at each sensor placed at various locations at time t. Given the incomplete feature values from each sensor at different time instances, the goal is to impute the missing feature values by considering spatial and temporal dependencies. The observability of the feature values/measurements is captured by a binary mask matrix $\mathbf{M}_t \in \{0,1\}^{N\times D}$ with elements as 1 if $\mathbf{X}_t^{(i,j)}$ is observed and 0 otherwise. Given the incomplete sensor measurements $\mathbf{Y}_t = \mathbf{M}_t \odot \mathbf{X}_t$, the objective is to reconstruct the missing sensor measurements and estimate the complete matrix $\hat{\mathbf{Y}}_t$. Analysis of spatiotemporal data becomes harder due to complicated underlying patterns. The use of a network topology aids in explicitly modeling and capturing the underlying complex spatiotemporal connections.

III. Proposed G-LSTM Framework

The proposed spatiotemporal imputation framework is depicted in figure 1. The framework consists of a spatial module integrated with a temporal module to estimate $\hat{\mathbf{Y}}_t$. These modules are described next.

Spatial Module: The spatial module in our framework captures the spatial relationships within the data. The module consists of multiple stacked layers of GraphSAGE (SAmple and aggreGatE) [14]. Each node in the graph is represented by a node embedding vector that helps capture the structural information. GraphSAGE is an inductive approach in which the algorithm learns a mapping (aggregator) function instead of the node embedding vector. The two primary steps associated with GraphSAGE are aggregate and update. Each node in the graph is uniquely represented by a feature vector. The aggregate step aggregates the neighboring node representations (feature vectors) for our target node. After obtaining an aggregated representation for node v based on its neighbors, the feature representation of the current node v is updated using a combination of its previous representation. GraphSAGE's inductive nature allows it to infer the node embedding vector for nodes not encountered during training, making it suitable for our imputation framework. The GNN module takes the spatial adjacency matrix and the observed feature values at each time instance. It captures the spatial information using the message-passing mechanism of a Graph Neural Network. The GNN output, which consists of features learned from spatial dependencies, is then passed on to the temporal module described next.

Temporal Module: Temporal modules are used for capturing the temporal relations in the data. In this work, the temporal module is driven by Long Short Term Memory (LSTM) memory networks [15] for our imputation framework. LSTMs have been shown to perform well on sequence-based tasks with long-term dependencies, assisting in capturing temporal associations. The reconstruction loss calculated during the training is on the entire observation and helps to learn the inherent relationships among the entire spatiotemporal dataset. The patterns learned and propagated help update/ impute data at the missing location and time instants.

The spatiotemporal data is loaded and punctured to create missing values (the missing locations are given zero values). Then, the data is split temporally into training and testing segments. Finally, the ST block is trained and the recovered values are used for evaluation by comparison with the original input data. The loss function used for evaluation corresponds to the following:

$$\mathbb{L}(\theta) = \sqrt{\sum_{v=1}^{N}\sum_{t=1}^{T}(x_t^v - \hat{x}_t^v)^2} \qquad (1)$$

where $\mathbb{L}(\theta)$ is the reconstruction error on both observed and missing data. x_t^v and $\hat{x}_t^v$ are the actual and estimated values respectively for node v and time t.

A. Complexity Analysis:

In this subsection, we discuss the computational complexities of the proposed algorithm. The computing cost of the G-LSTM is concentrated on the GraphSAGE layers and the LSTM layers. Suppose for a graph that is being considered there are n number of nodes, r number of neighbors being sampled for each node and m the total number of edges and K number of layers. The computational complexity of the

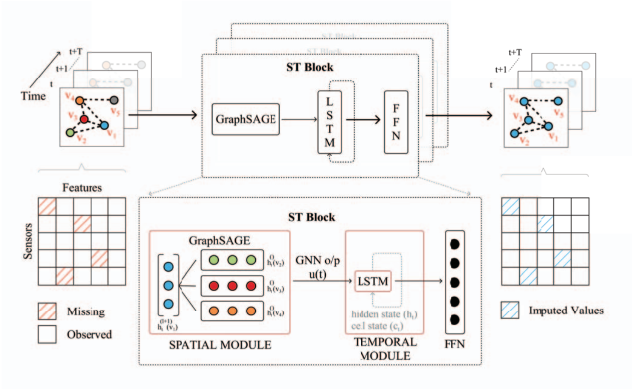

Fig. 1. Structure of proposed spatiotemporal imputation network

GraphSAGE can be represented as $\mathcal{O}(r^K nd^2)$. For an LSTM with h as the hidden layer dimension and b as the input feature, the computational complexity is $\mathcal{O}(4h(b+2+h))$. Thus the computational complexity of the G-LSTM framework can be approximated as $\mathcal{O}(r^K nd^2 + 4h(b+2+h))$.

IV. Results

In this section, we evaluate the performance of the proposed imputation method on a real-world traffic dataset. We introduce the dataset and explain the experimental settings first. Then various missing data scenarios are considered. The imputation performance of the proposed method is compared with Costco, a neural tensor completion-based imputation framework as the baseline.

A. Traffic-State dataset:

PeMSD7 is a traffic dataset collected from Caltrans Performance Measurement System (PeMS) and describes the speed detectors covering the freeway system in all major California urban centers. Among the 1000 sensors placed on the arterial roads of District 7 in California, we chose 228 sensors for our study, similar to [16]. The dataset consists of 5-min average traffic speed data collected from May 1, 2012, to June 30, 2012, and has 11232 time points.

The graphical representation of the PEMS dataset was created in a manner similar to [16]. The nodes correspond to each sensor in the network and the initial node features are the speed values concatenated with the singular value decomposition (SVD) of the weighted adjacency matrix. The SVD values concatenated with the speed values aid in uniquely identifying the nodes [17]. For computational purposes, the data is normalized. The imputation performance is evaluated using two metrics, root mean square error (RMSE) and mean absolute error (MAE) defined as,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \widehat{x_i}| \tag{3}$$

where, n is the number of missing values, x_i is the i^{th} missing value and $\widehat{x_i}$ is the imputed value of x_i.

B. Experimental Results:

For training and testing purposes, the dataset is split temporally. The initial 80% of time instances are used for training purposes and the remaining 20% is used for testing. The data is scaled to be between 0 and 1 before imputation and the output of the proposed approach is rescaled back to the original values. It is assumed that the sensors at specific locations were absent across all time instances to simulate the missing values.

The proposed imputation framework model is trained by Adam optimizer [18]. The learning rate is set at 0.001 with the ReLu activation function after the GNN layers. The time window for the LSTM is set as one in order to maintain an end-to-end training framework and the temporal information is captured by the hidden and the cell states. All experiments are implemented with PyTorch and conducted on an NVIDIA GeForce RTX 3070 GPU.

To evaluate the performance of the proposed approach, we compare its performance with CoSTCo [7], a state-of-the-art convolution neural network-based tensor completion approach. CoSTCo tries to address the inability of multilinear models to generate a low-rank representation of non-linear data. It utilizes the activation functions in a convolutional neural network to capture non-linear relationships. Table I summarizes the experimental results using our proposed approach compared with CoSTCo as the baseline. The imputation errors are calculated with various missing ratios for a single time instant. The results are averaged over five random missing scenarios for each of the missing ratios. It can be seen that the proposed approach (G-LSTM) outperforms the baseline and as the percentage of missing information increases, the imputation error metrics remain steady and sometimes show improvement. This surprising observation can be attributed to the possibility that at a lower percentage of missing data, the GNN may be overfitting. As fewer data are available (or the missing percentage increases), the overfitting behavior is less likely. The G-LSTM is trained with 50% of missing data and the pre-trained model is used for imputing the data with 10%, 30%, 70% and 90% of missing information. From the results shown in table I, it can be seen that for missing percentages lesser than 50%, the pre-trained G-LSTM model gives improved results compared to the corresponding G-LSTM model and CoSTCo. For missing percentages greater than 50%, the pre-trained G-LSTM outperforms CoSTCo. However, the pre-trained model shows slightly inferior performance with respect to G-LSTM. The results indicate that the model trained for higher missing percentages can impute the data with lower missing information more efficiently and thus possess the

TABLE I
ERROR METRICS FOR MISSING DATA IMPUTATION

% Missing	10 %		30%		50%		70%		90%	
Model	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
G-LSTM	1.61	1.90	1.67	2.09	0.74	1.19	1.16	1.50	0.83	1.23
CoSTCo	2.26	2.27	2.60	2.61	4.66	4.67	6.96	6.97	7.90	7.91
G-LSTM (pre-trained)	0.72	1.36	0.92	1.23	NA	NA	1.33	1.59	1.65	1.85

inductive capability. Relative to the baseline CoSTCo method, the proposed framework offers approximately 69% reduced MAE and 61% reduced RMSE on the imputed values.

The speed values at each sensor for a single time step as predicted by G-LSTM and CoSTCo when 50% of information is missing are shown in figures 2 and 3, respectively. It can be seen that our framework outperforms the baseline in prediction with missing information.

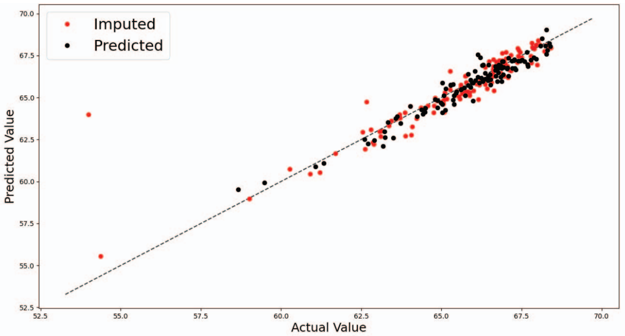

Fig. 2. Speed values estimated for each node for a single time instant using G-LSTM framework

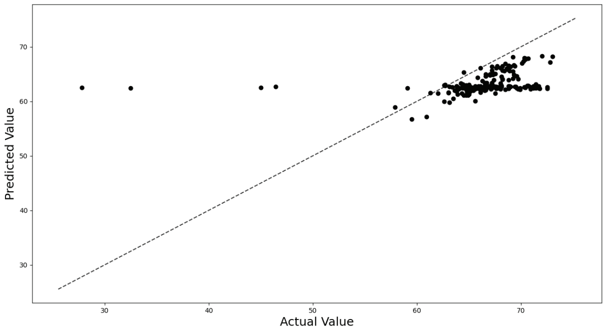

Fig. 3. Speed values estimated for each node for a single time instant using CosTCo

V. Conclusions and Future Work

In this paper, we propose a general spatiotemporal framework for data imputation. The traffic sensor network is formulated as an undirected graph with speed values as a time-varying feature. The proposed imputation framework consists of a spatial and a temporal module that helps capture the spatiotemporal relationships within the data and thereby helps to impute the missing information. Comprehensive case studies are conducted to evaluate the imputation accuracy of the proposed model for a wide missing rate range. Experimental results show that the proposed method outperforms the neural tensor completion method, CoSTCo and maintains steady performance in extreme missing scenarios. In future studies, we plan to improve imputation accuracy by employing an attention mechanism instead of LSTM and incorporating model information instead of a purely data-driven approach.

Acknowledgement

The research was funded by National Science Foundation (NSF) CNS 2039014

References

[1] S. R. Kuppannagari, Y. Fu, C. M. Chueng, and V. K. Prasanna, "Spatio-temporal missing data imputation for smart power grids," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, 2021, pp. 458–465.

[2] B. Bae, H. Kim, H. Lim, Y. Liu, L. D. Han, and P. B. Freeze, "Missing data imputation for traffic flow speed using spatio-temporal cokriging," *Transportation Research Part C: Emerging Technologies*, vol. 88, pp. 124–139, 2018.

[3] H. Yang, J. Yang, L. D. Han, X. Liu, L. Pu, S.-m. Chin, and H.-l. Hwang, "A kriging based spatiotemporal approach for traffic volume data imputation," *PloS one*, vol. 13, no. 4, p. e0195957, 2018.

[4] T. F. Johnson, N. J. Isaac, A. Paviolo, and M. González-Suárez, "Handling missing values in trait data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51–62, 2021.

[5] A. Cini, I. Marisca, and C. Alippi, "Filling the g_ap_s: Multivariate time series imputation by graph neural networks," *arXiv preprint arXiv:2108.00298*, 2021.

[6] X. Chen, M. Lei, N. Saunier, and L. Sun, "Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[7] H. Liu, Y. Li, M. Tsang, and Y. Liu, "Costco: A neural tensor completion model for sparse tensors," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 324–334.

[8] J. Poloczek, N. A. Treiber, and O. Kramer, "Knn regression as geo-imputation method for spatio-temporal wind data," in *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*. Springer, 2014, pp. 185–193.

[9] N. Marchang and R. Tripathi, "Knn-st: Exploiting spatio-temporal correlation for missing data inference in environmental crowd sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3429–3436, 2020.

[10] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transportation research part C: emerging technologies*, vol. 72, pp. 168–181, 2016.

[11] Y. Ye, S. Zhang, and J. J. Yu, "Spatial-temporal traffic data imputation via graph attention convolutional network," in *International Conference on Artificial Neural Networks*. Springer, 2021, pp. 241–252.

[12] Y. Wu, D. Zhuang, A. Labbe, and L. Sun, "Inductive graph neural networks for spatiotemporal kriging," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4478–4485.

[13] A. J. Saroj, A. Guin, and M. Hunter, "Deep lstm recurrent neural networks for arterial traffic volume data imputation," *Journal of Big Data Analytics in Transportation*, vol. 3, no. 2, pp. 95–108, 2021.

[14] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[15] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[16] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[17] V. Lingam, R. Ragesh, A. Iyer, and S. Sellamanickam, "Simple truncated svd based model for node classification on heterophilic graphs," *arXiv preprint arXiv:2106.12807*, 2021.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.