Mitigating missing data for Human Activity Recognition

Team:
Rushi Bhatt
Ronak Kaoshik
Jinru Cai

Overall Project Goals

- Building robust models for missing data imputation and downstream task of HAR classification

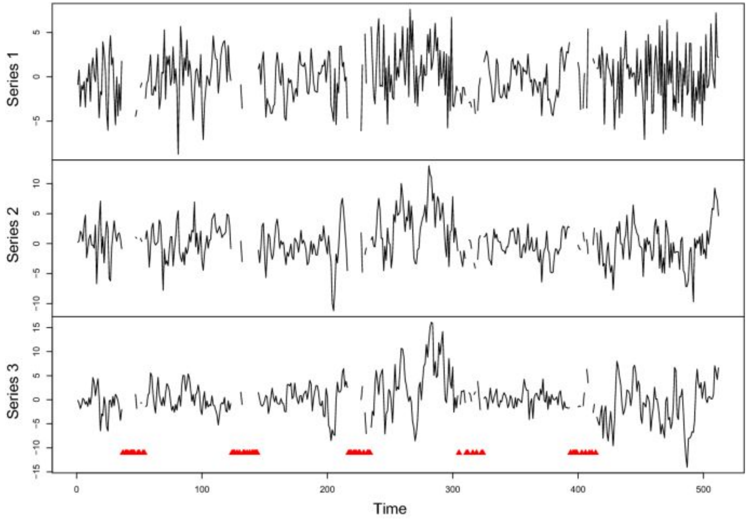


Image Source: https://link.springer.com/article/10.1007/s11222-021-09998-2

Specific Aims

- Investigation on types of missingness for HAR
- Extending SOTA imputation models for HAR
- Perform a comparative study between SOTA and baseline models for reconstruction and classification
  - MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and MRE (Mean Relative Error) for imputation techniques
  - ROC-AUC (Area Under ROC Curve), PR-AUC (Area Under Precision-Recall Curve), and F1-score for downstream classification
  - Across different missingness rates and types

Dataset

- UCI HAR [1]
  - 6 activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING)
  - 128 readings/ window (~10k time series)
  - 3-axis accelerometer and 3-axis gyroscope
- PAMAP 2 [2, 3]
  - 24 activities
  - 52 features
    - Heart Beat
    - IMU hand (2 x 3-axis accelo, 3-axis gyro, 3-axis magneto, orientation)
    - IMU chest
    - IMU ankle
- Real-life dataset

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013

[2] A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

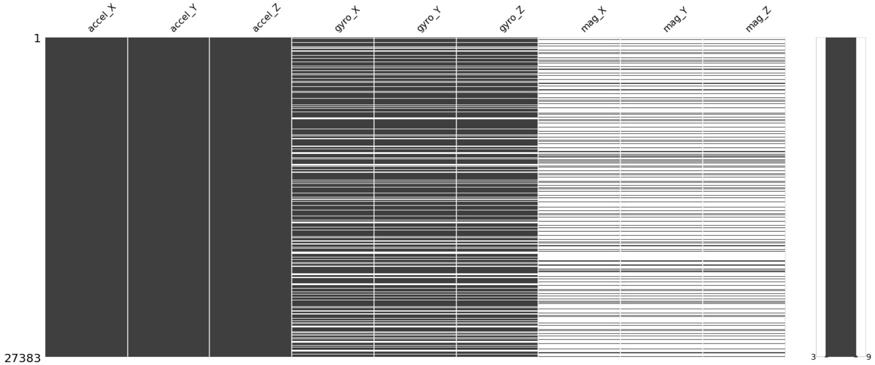
[3] A. Reiss and D. Stricker. Creating and Benchmarking a New Dataset for Physical Activity Monitoring. The 5th Workshop on Affect and Behaviour Related Assistance (ABRA), 2012.

Data Missingness

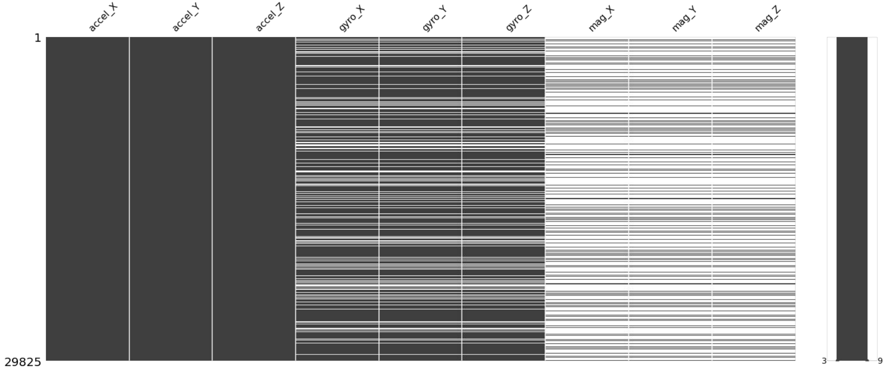
Types of data missingness:

- Missing completely at random (MCAR), missing values are independent of any other values
- Missing at random (MAR), missing values depend only on observed values
    - E.g., missingness rate for walking is higher than that for standing?
    - E.g., missingness rate for sensor X is higher than sensor Y?
- Missing not at random (MNAR), missing values depend on both observed and unobserved values
    - E.g., if you run very fast (more acceleration), missingness is likely to be higher than normal run

Data Missingness



(a) Missingness matrix for standing class



(b) Missingness matrix for walking class

	Standing Class	Walking Class
AccelX, AccelY, AccelZ	0.00 %	0.00 %
GyroX, GyroY, GyroZ	2.76 %	2.51 %
MagX, MagY, MagZ	8.83 %	8.77 %

Table: Class-wise % missingness comparison

*1 min of walking and standing data recorded using Sensorstream IMU+GPS app for Android

Current approaches

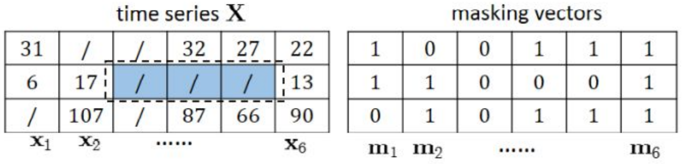
Baseline imputation techniques

- **Mean**: It involves simply replacing the missing values with corresponding global mean.
- **KNN**: It uses k-nearest neighbor (with normalized Euclidean distance) to find the similar samples, and impute the missing values with weighted average of its neighbors.
- **Matrix Factorization (MF)**: It factorizes the data matrix into two low-rank matrices, and fill the missing values by matrix completion.
- **MICE**: It uses Multiple Imputation by Chained Equations (MICE), a widely used imputation method, to fill the missing values. MICE creates multiple imputations with chained equations.
- **ImputeTS**: It uses ImputeTS package in R to impute the missing values. ImputeTS is a widely used package for missing value imputation, which utilizes the state space model and kalman smoothing.

[4] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, ‘BRITS: Bidirectional Recurrent Imputation for Time Series’, στο Advances in Neural Information Processing Systems, 2018, τ. 31.

BRITS[4]:Bidirectional Recurrent Imputation for Time Series

Unidirectional Uncorrelated Recurrent Imputation RITS-I



time series $\mathbf{X}$

31	/	/	32	27	22
6	17	/	/	/	13
/	107	/	87	66	90

$\mathbf{x}_1$ $\mathbf{x}_2$ ...... $\mathbf{x}_6$

masking vectors

1	0	0	1	1	1
1	1	0	0	0	1
0	1	0	1	1	1

$\mathbf{m}_1$ $\mathbf{m}_2$ ...... $\mathbf{m}_6$

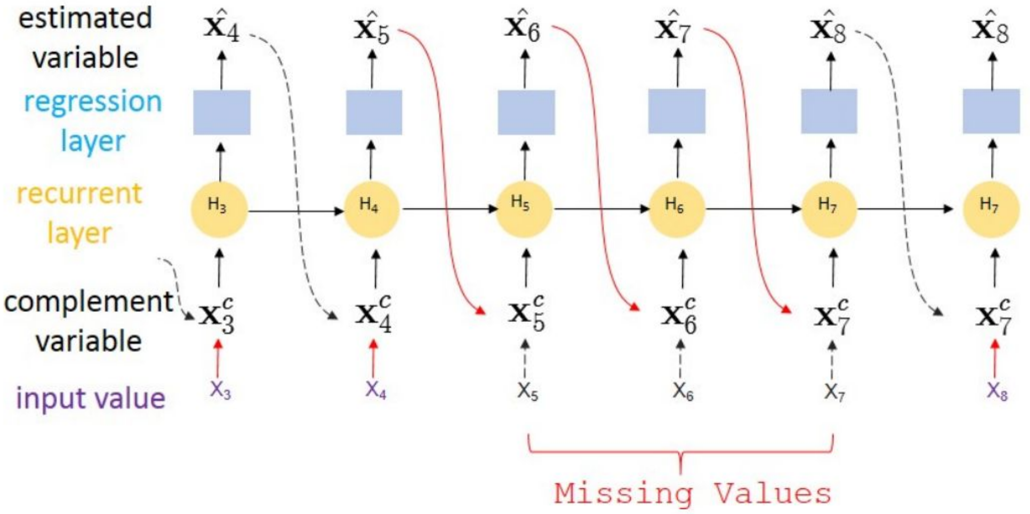


Figure 2: Imputation with unidirectional dynamics.

[4] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, ‘BRITS: Bidirectional Recurrent Imputation for Time Series’, στο Advances in Neural Information Processing Systems, 2018, τ. 31.

BRITS[4]:Bidirectional Recurrent Imputation for Time Series

Vanilla RNN function: $\mathbf{h}_t = \sigma(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{U}_h \mathbf{x}_t + \mathbf{b}_h),$

$$\hat{\mathbf{x}}_t = \mathbf{W}_x \mathbf{h}_{t-1} + \mathbf{b}_x, \tag{1}$$

$$\mathbf{x}_t^c = \mathbf{m}_t \odot \mathbf{x}_t + (1 - \mathbf{m}_t) \odot \hat{\mathbf{x}}_t, \tag{2}$$

$$\gamma_t = \exp\{-\max(0, \mathbf{W}_\gamma \delta_t + \mathbf{b}_\gamma)\}, \tag{3}$$

$$\mathbf{h}_t = \sigma(\mathbf{W}_h[\mathbf{h}_{t-1} \odot \gamma_t] + \mathbf{U}_h[\mathbf{x}_t^c \circ \mathbf{m}_t] + \mathbf{b}_h), \tag{4}$$

$$\ell_t = \langle \mathbf{m}_t, \mathcal{L}_e(\mathbf{x}_t, \hat{\mathbf{x}}_t) \rangle\,. \tag{5}$$

$$\hat{\mathbf{y}} = f_{out}(\sum_{i=1}^{T} \alpha_i \mathbf{h}_i), \quad \frac{1}{T}\sum_{t=1}^{T} \ell_t + \mathcal{L}_{out}(\mathbf{y}, \hat{\mathbf{y}}) \quad \mathtt{MAE} = \frac{\sum_i |\mathtt{pred}_i - \mathtt{label}_i|}{N}, \quad \mathtt{MRE} = \frac{\sum_i |\mathtt{pred}_i - \mathtt{label}_i|}{\sum_i |\mathtt{label}_i|}.$$

Loss function

[4] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, ‘BRITS: Bidirectional Recurrent Imputation for Time Series’, στο Advances in Neural Information Processing Systems, 2018, τ. 31.

BRITS[4]:Bidirectional Recurrent Imputation for Time Series

Bidirectional Uncorrelated Recurrent Imputation BRITS-I

$$\ell_t^{cons} = \text{Discrepancy}(\hat{\mathbf{x}}_t, \hat{\mathbf{x}}_t')$$

Unidirectional correlated Recurrent Imputation RITS

$$\hat{\mathbf{z}}_t = \mathbf{W}_z \mathbf{x}_t^c + \mathbf{b}_z,$$

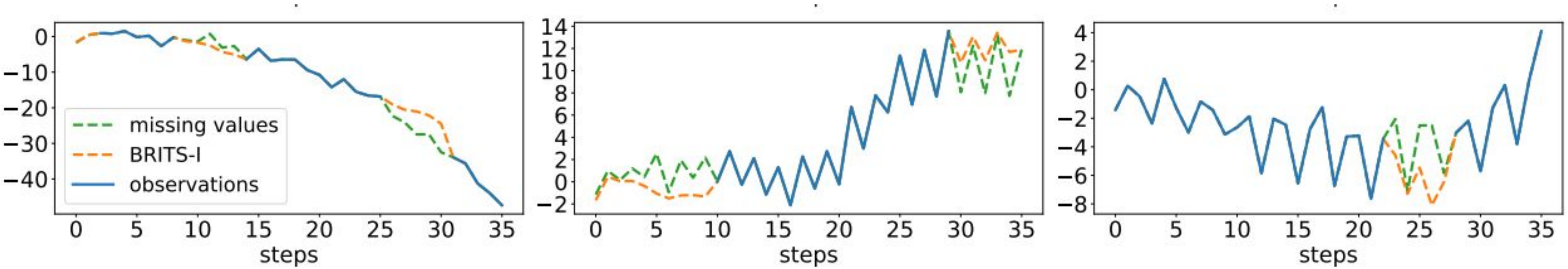
$$\begin{aligned} \beta_t &= \sigma(\mathbf{W}_\beta[\gamma_t \circ \mathbf{m}_t] + \mathbf{b}_\beta) \\ \hat{\mathbf{c}}_t &= \beta_{\mathbf{t}} \odot \hat{\mathbf{z}}_t + (1 - \beta_{\mathbf{t}}) \odot \hat{\mathbf{x}}_t. \end{aligned}$$

$$\begin{aligned} \mathbf{c}_t^c &= \mathbf{m}_t \odot \mathbf{x}_t + (1 - \mathbf{m}_t) \odot \hat{\mathbf{c}}_t \\ \mathbf{h}_t &= \sigma(\mathbf{W}_h[\mathbf{h}_{t-1} \odot \gamma_t] + \mathbf{U}_h[\mathbf{c}_t^c \circ \mathbf{m}_t] + \mathbf{b}_h). \end{aligned}$$

$$\ell_t = \mathcal{L}_e(\mathbf{x}_t, \hat{\mathbf{x}}_t) + \mathcal{L}_e(\mathbf{x}_t, \hat{\mathbf{z}}_t) + \mathcal{L}_e(\mathbf{x}_t, \hat{\mathbf{c}}_t).$$

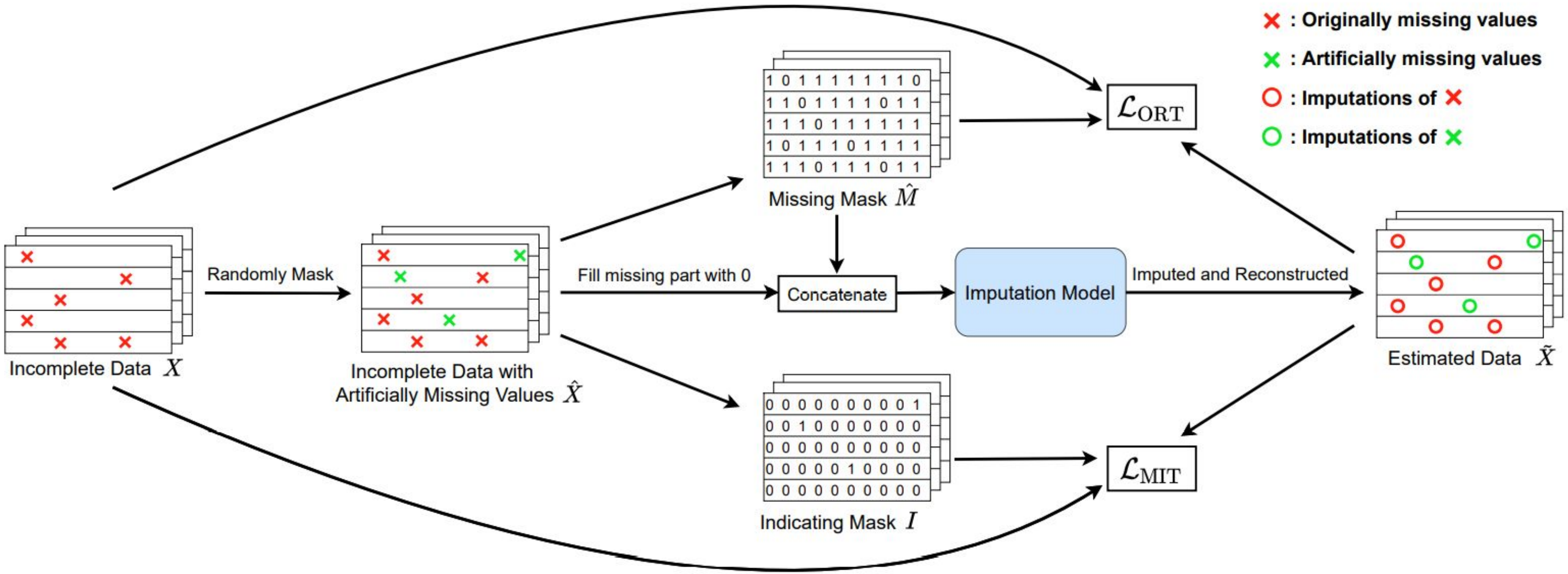
[4] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, ‘BRITS: Bidirectional Recurrent Imputation for Time Series’, στο Advances in Neural Information Processing Systems, 2018, τ. 31.

BRITS[4]:Bidirectional Recurrent Imputation for Time Series



[4] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, ‘BRITS: Bidirectional Recurrent Imputation for Time Series’, στο Advances in Neural Information Processing Systems, 2018, τ. 31.

SAITS [5]: Self-Attention-Based Imputation For Time Series


✕ : Originally missing values
✕ : Artificially missing values
○ : Imputations of ✕
○ : Imputations of ✕
Incomplete Data X
Randomly Mask
Incomplete Data with Artificially Missing Values $\hat{X}$
Fill missing part with 0
Missing Mask $\hat{M}$
Concatenate
Imputation Model
Imputed and Reconstructed
Estimated Data $\tilde{X}$
Indicating Mask I
$\mathcal{L}_{\text{ORT}}$
$\mathcal{L}_{\text{MIT}}$

Missing Mask $\hat{M}$

1	0	1	1	1	1	1	1	1	0
1	1	0	1	1	1	1	0	1	1
1	1	1	0	1	1	1	1	1	1
1	0	1	1	1	0	1	1	1	1
1	1	1	0	1	1	1	0	1	1

Indicating Mask I

0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Figure 1: A graphical overview of the joint-optimization training approach.

[5] W. Du, D. Côté, και Y. Liu, ‘SAITS: Self-Attention-based Imputation for Time Series’. arXiv, 2022.

SAITS [5]: Self-Attention-Based Imputation For Time Series

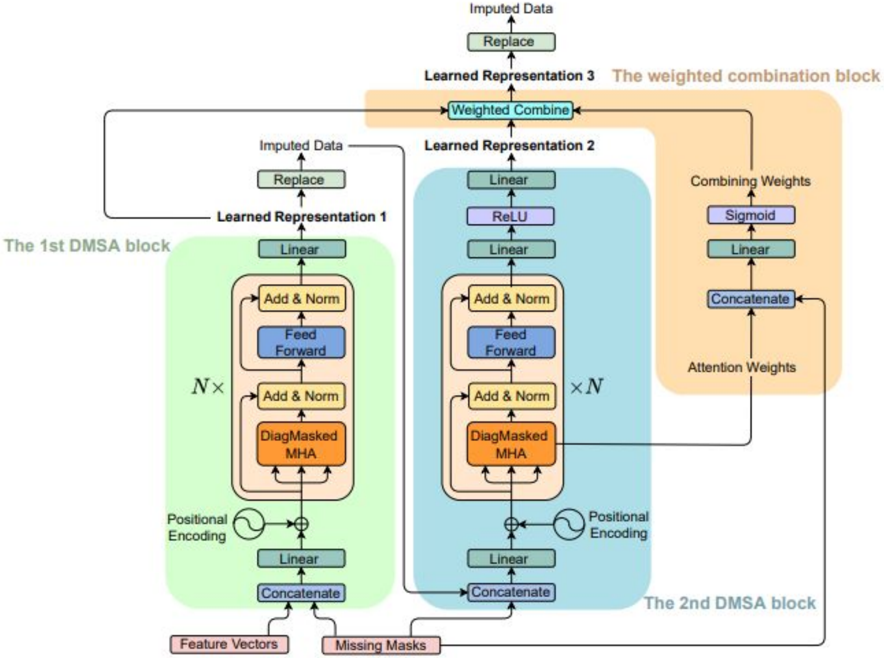
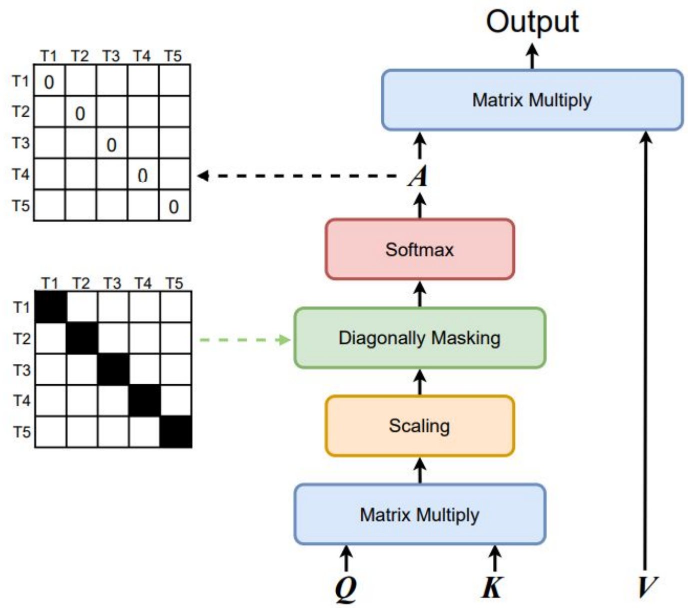


Figure 3: The SAITS model architecture.

[5] W. Du, D. Côté, και Y. Liu, 'SAITS: Self-Attention-based Imputation for Time Series'. arXiv, 2022.

SAITS [5]: Self-Attention-Based Imputation For Time Series



	T1	T2	T3	T4	T5
T1	0				
T2		0			
T3			0		
T4				0	
T5					0

	T1	T2	T3	T4	T5
T1	■				
T2		■			
T3			■		
T4				■	
T5					■

Diagonally-masked self-attention on a time-series sample with five time steps.

[5] W. Du, D. Côté, και Y. Liu, ‘SAITS: Self-Attention-based Imputation for Time Series’. arXiv, 2022.

SAITS [5]: Self-Attention-Based Imputation For Time Series

$$e = \left[\mathrm{Concat}\left(\hat{X}, \hat{M}\right) W_e + b_e\right] + p$$

$$z = \{\mathrm{FFN}(\mathrm{DiagMaskedMHA}\,(e))\}^N$$

$$\tilde{X}_1 = zW_z + b_z$$

$$\hat{X}' = \hat{M} \odot \hat{X} + \left(1 - \hat{M}\right) \odot \tilde{X}_1$$

First DMSA block

$$\alpha = \left[\mathrm{Concat}\left(\hat{X}', \hat{M}\right) W_\alpha + b_\alpha\right] + p$$

$$\beta = \{\mathrm{FFN}\,(\mathrm{DiagMaskedMHA}\,(\alpha))\}^N$$

$$\tilde{X}_2 = \mathrm{ReLU}\,(\beta W_\beta + b_\beta)\, W_\gamma + b_\gamma$$

Second DMSA block

$$\mathcal{L}_{\mathrm{ORT}} = \frac{1}{3}\left(\ell_{\mathrm{MAE}}\left(\tilde{X}_1, X, \hat{M}\right) + \ell_{\mathrm{MAE}}\left(\tilde{X}_2, X, \hat{M}\right) + \ell_{\mathrm{MAE}}\left(\tilde{X}_3, X, \hat{M}\right)\right)$$

$$\mathcal{L}_{\mathrm{MIT}} = \ell_{\mathrm{MAE}}\left(\hat{X}_c, X, I\right)$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{ORT}} + \lambda\, \mathcal{L}_{\mathrm{MIT}}$$

[5] W. Du, D. Côté, και Y. Liu, ‘SAITS: Self-Attention-based Imputation for Time Series’. arXiv, 2022.

Evaluation metrics

Imputation task

$$\mathrm{MAE}\left(estimation, target, mask\right) = \frac{\sum_d^D \sum_t^T |(estimation - target) \odot mask|_t^d}{\sum_d^D \sum_t^T mask_t^d}$$

$$\mathrm{RMSE}\left(estimation, target, mask\right) = \sqrt{\frac{\sum_d^D \sum_t^T \left(((estimation - target) \odot mask)^2\right)_t^d}{\sum_d^D \sum_t^T mask_t^d}}$$

$$\mathrm{MRE}\left(estimation, target, mask\right) = \frac{\sum_d^D \sum_t^T |(estimation - target) \odot mask|_t^d}{\sum_d^D \sum_t^T |target \odot mask|_t^d}$$

Classification task

Accuracy, F1-Score, Confusion matrix, ROC-AUC

Related Work

Methods	Model description	Dataset	Pros	Cons
Bidirectional Recurrent Imputation for Time Series (BRITS)	novel method based on recurrent neural networks for missing value imputation in time series data. method directly learns the missing values in a bidirectional recurrent dynamical system.	three real-world datasets, including an air quality dataset, a health-care dataset and a localization dataset of human activities.	BRITS is robust to multiple correlated missing values and can be applied to different settings(datasets) as a data-driven imputation procedure.	BRITS has eliminated just 10% of the time-series data randomly from the ground truth.
Self-Attention-Based Imputation For Time Series (SAITS)	missing values from a weighted combination of two diagonally-masked self-attention blocks. captures both the temporal dependencies and feature correlations between time steps	Beijing Multi-Site Air-Quality Dataset 7 papers also use this dataset	which explicitly capture both the temporal dependencies and feature correlations for between time steps which in turn improves imputation accuracy and training speed.	However, the reconstruction accuracy is not up to the mark

Method	Model Description	dataset	Pros	Cons
Directly Modeling Missing Data in Sequences with RNNs [7]	used a GRU-based network for the clinical time-series classification with missing data	dataset consists of patient records extracted from the EHR system at CHLA (Marlin et al., 2012; Che et al., 2015)	The structure of model is not very complex.	have memory constraints when dealing with long-term dependency in time series when the number of time-steps missing in the data sample is relatively big
E2GAN: End-to-End Generative Adversarial Network for Multivariate Time Series Imputation [8]	a generative adversarial network-based architecture has been introduced for the imputation task	PhysioNet Challenge 2012 dataset (PhysioNet). , is made up of records from 4000 ICU stays.	solves the imputation task in a single stage while making use of an auto-encoder based GRU network in the generator block.	involve a complex training cycle caused due to issues such as non-convergence and mode-collapse due to their respective loss formulations.
GP-VAE: Deep Probabilistic Time Series Imputation [9]	a variational auto-encoder (VAE) architecture for the imputation of time series along with a Gaussian process (GP) prior defined in the latent space	SPRITES data set consists of 9,000 sequences of animated characters with different clothes, hair styles, and skin colors, performing different actions,	This GP-prior is helps with the embedding of data into a smooth explainable representation.	involve a complex training cycle

[7] Z. C. Lipton et al., “Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series,” Mach. Learning Healthcare conf., pp. 253–270, 2016.

[8] Yonghong Luo, Ying Zhang, Xiangrui Cai, and Xiaojie Yuan. E2GAN: End-to-end generative adversarial network for multivariate time series imputation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 3094–3100. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[9] Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In Silvia Chiappa and Roberto Calandra, editors, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 1651–1661. PMLR, 26–28 Aug 2020.

Current Status

- Performed extensive literature survey on SOTA imputation methods
- Started investigating on data missingness distribution
- Replicating BRITS implementation

Progress up to date in GitHub Repo:

- https://github.com/RushiBhatt007/ece209as_project

Next Steps

- Deeper analysis on selection of missingness type
- Implementing baseline imputation and classification models for HAR
- Replicating BRITS and SAITS for HAR
- Performance comparison between SOTA and baseline approaches
- Comparative study for multiple missingness types and missingness thresholds

Thank you