End-to-end Multi-task Learning of Missing Value Imputation and Classification in Time-series Data

Jinhee Kim[1*] Taesung Kim[1*] Jang-Ho Choi[2] Jaegul Choo[1]

[1]KAIST, Daejeon, South Korea
[2]Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

*Equal contribution

Contents

1 Introduction

- Multivariate time-series prediction often becomes challenging due to missing data caused by unreliable sensors and other issues.
- Inaccurate imputation of missing values can degrade the downstream prediction performance.
  - It may be better not to rely on the estimated values of missing data.
  - Observed data may contain noise.

- We propose a novel approach that can **automatically utilize the optimal combination of the observed and the estimated values** to generate not only complete, but also noise-reduced data by our own gating mechanism.
- **By jointly training the downstream task module and gating mechanism with adversarial loss**, our model produces realistic and helpful imputation to predict the downstream task.
- We also **design a synthetic dataset with a known true distribution** to verify our method.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

Imputation

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Downstream task

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

classification/regression

Mortality/ PM10, ...

2 Proposed method

Additional dropping of the input data

Colors indicate the status of each variable:

- *white* for observed variables,
- *red* for missing variables, and
- *blue* for **additionally dropped variables.**

X: raw data matrix

x_1	x_2	x_3	x_4	x_5
-20	17	-50	91	25
NaN	NaN	26	NaN	64
37	NaN	13	NaN	NaN

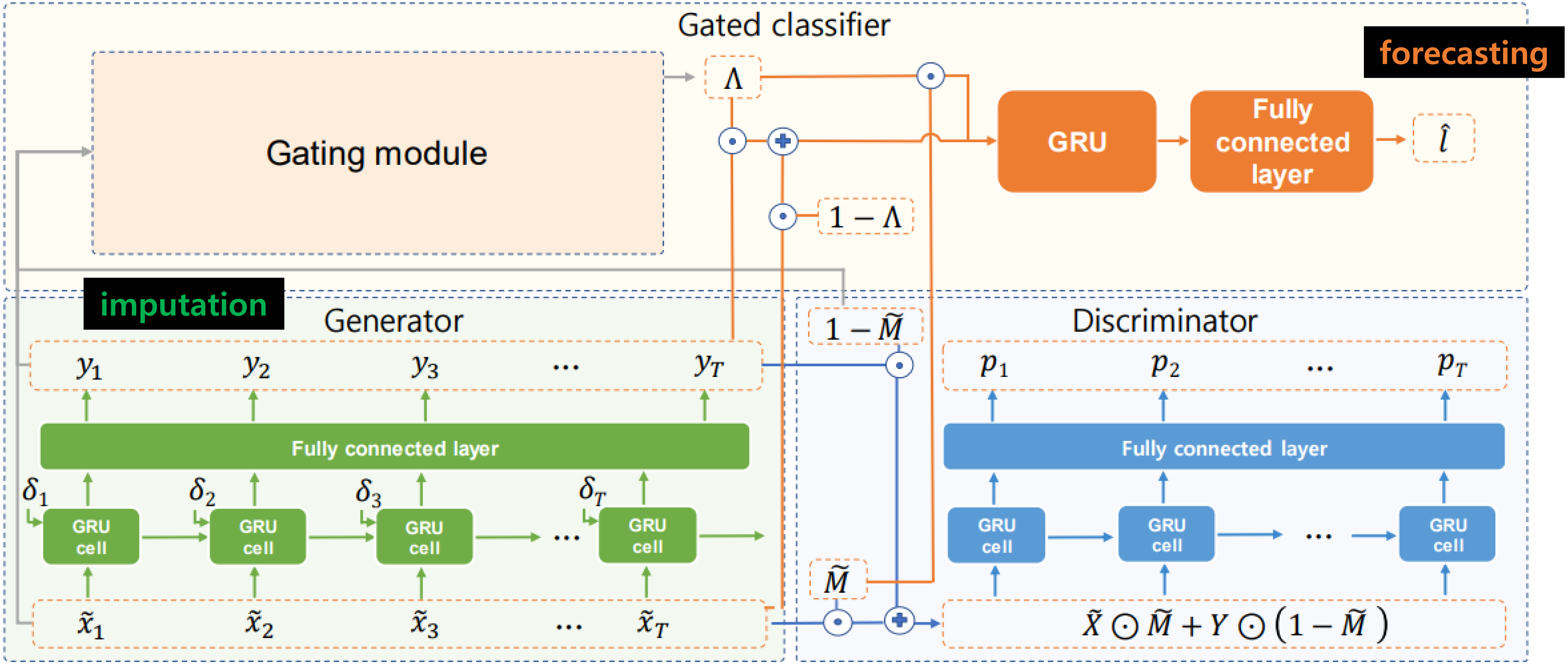
Input to the network

$\tilde{X}$: dropped data matrix

$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$
NaN	17	-50	91	NaN
NaN	NaN	26	NaN	64
37	NaN	NaN	NaN	NaN

2 Proposed method

An end-to-end multitask learning of missing value imputation and forecasting



GRU cell with decaying mechanism for missing values

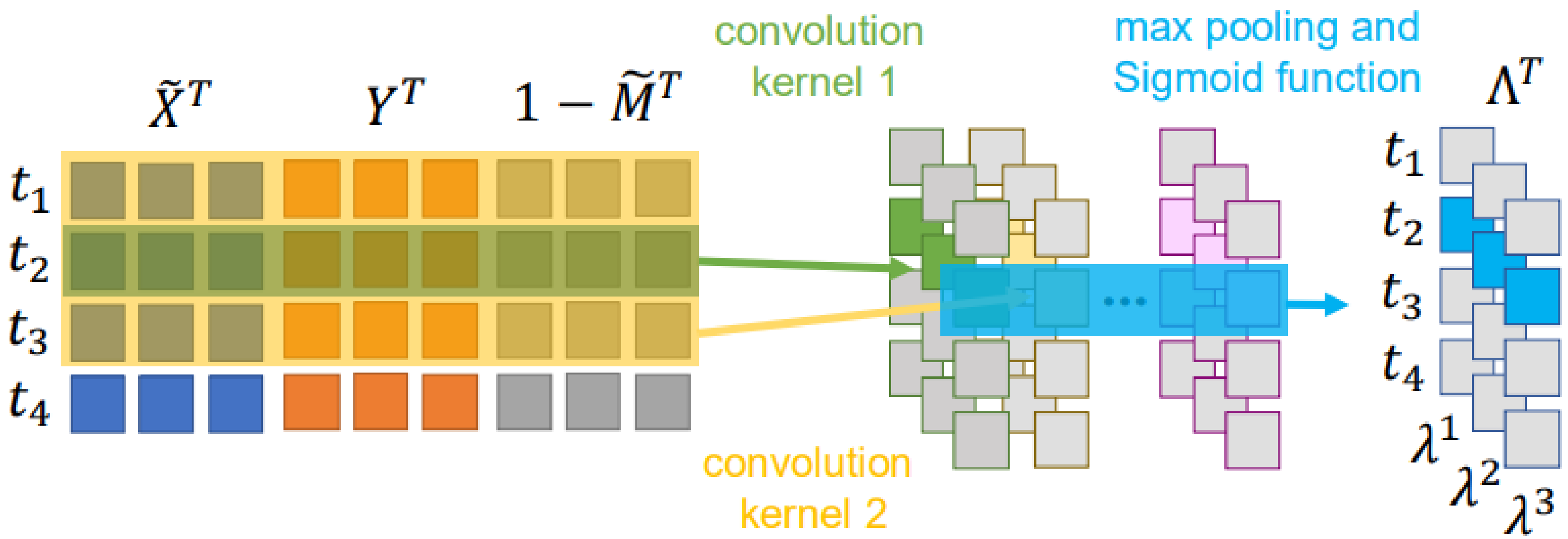
- The missing pattern of a variable with respect to time should be considered.
- Thus, we propose to **decay the hidden state vector of GRU cell if a variable has been missing for a long while:**

$$\eta_t = max(0, W_\eta \delta_{t-k+1:t} + b_\eta)$$
$$\gamma_t = exp(-max(Maxpool(0, W_\gamma \eta_t^T + b_\gamma)^T))$$

- The **update functions of GRU** with the decaying mechanism is as follows:

$$h_{t-1} = \gamma_t \odot h_{t-1}, \quad z_t = [(1-\gamma_t); 0] \odot z_t$$
$$u_t = \sigma(W_u[h_{t-1}; z_t] + b_u), \quad r_t = \sigma(W_r[h_{t-1}; z_t] + b_r)$$
$$\tilde{h}_t = tanh(W_h[r_t \odot h_{t-1}; z_t] + b_h)$$
$$h_t = (1-u_t) \odot h_{t-1} + u_t \odot \tilde{h}_t$$

Gating module for down stream classification task



The generator output Y and the raw data $\tilde{X}$ are mixed by the ratio of gating value, Λ.

$$S = Y \odot \Lambda + \tilde{X} \odot (1 - \Lambda)$$

Since there is a shortfall in the scale of the missing values compared to the observed values, the mixed output S is compensated with the GRU weights.
The combination of the observed and the estimated values is fed to the classifier.

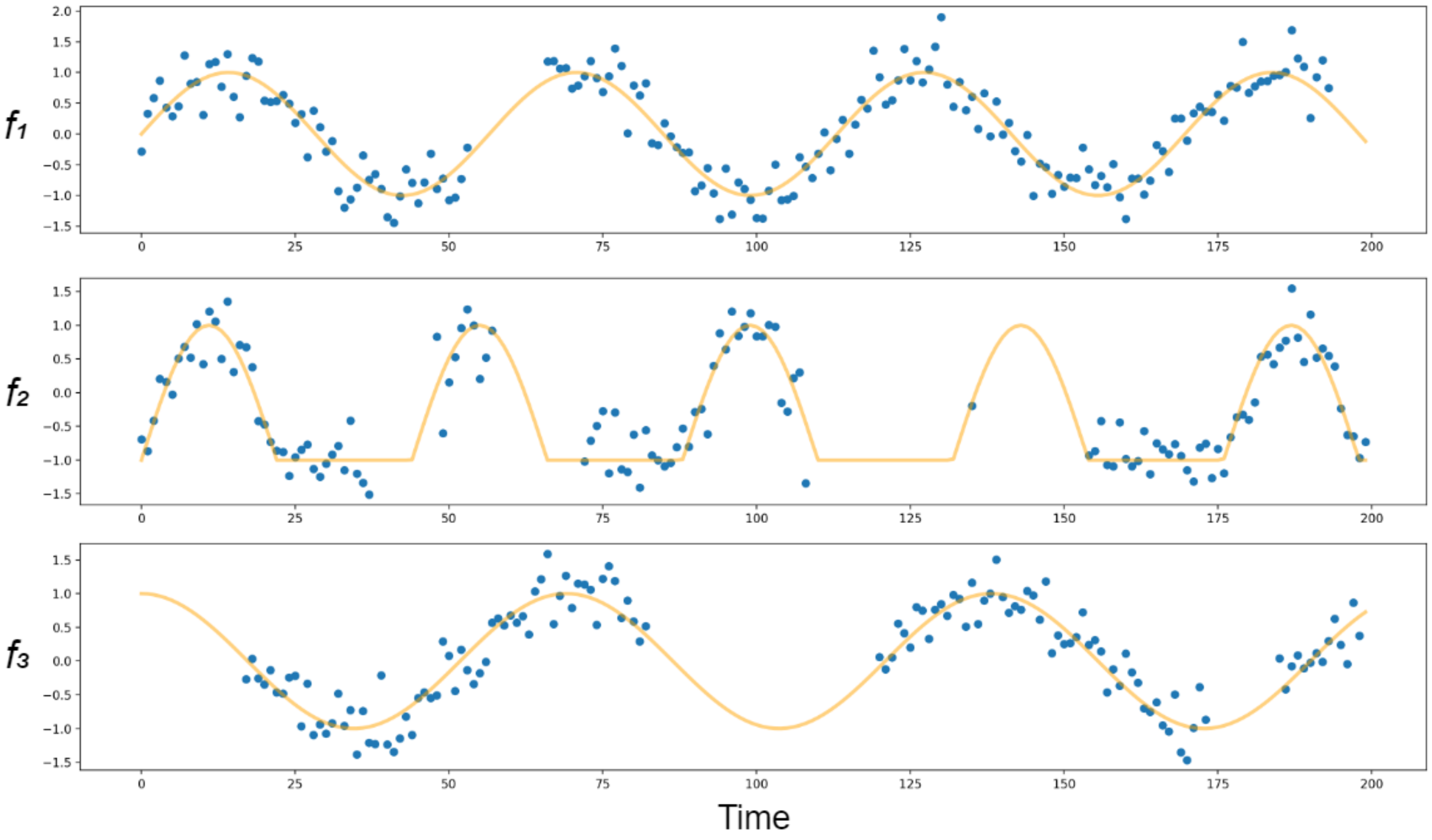
3 Synthetic Dataset

- A toy multivariate time-series dataset designed for missing value imputation in time-series data.
- A model can be validated with completely known data distributions.
- Data are randomly removed or added with Gaussian noise.
- The classification label is set to be one i f the first and the second features with out noise are both positive. Otherwise, the label is zero.

$$f_1(t) = \sin\frac{t}{9} + \epsilon$$

$$f_2(t) = \sin\frac{t}{7} + \left|\ \sin\frac{t}{7}\ \right| - 1 + \epsilon$$

$$f_3(t) = \cos\frac{t}{11} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 0.3^2)$$



4 Experimental results

- **Imputation accuracy on PhysioNet Challenge 2012 dataset**
  - We randomly discard 10% of the validation set to measure the imputation performance.
  - Every experiments are conducted five times. The average value and standard deviation (in parentheses) of performances are reported.

  - Our proposed method outperforms the other imputation methods.

TABLE I
IMPUTATION PERFORMANCES ON THE PHYSIONET DATASET.

Method	MSE	MAE
zero imputation	1.051 (0.110)	0.706 (0.007)
mean imputation	0.830 (0.007)	0.459 (0.004)
LOCF	0.727 (0.073)	0.458 (0.002)
GRU-D	1.258 (0.061)	0.824 (0.015)
GRUI	1.329 (0.909)	0.840 (0.041)
E^2GAN	0.756 (0.100)	0.555 (0.009)
BRITS	1.022 (0.110)	0.632 (0.006)
ours	**0.477 (0.065)**	**0.382 (0.239)**

4 Experimental results

- **Mortality prediction performances on PhysioNet Challenge 2012 dataset**
  - For the models only designed for imputation without a classifier, an additional training step on the downstream classification task is conducted after imputation.
  - Our model performed better than other baselines, suggesting that the proposed model creates an imputation result helpful in predicting the downstream task.

TABLE II
MORTALITY CLASSIFICATION PERFORMANCES ON PHYSIONET DATASET. AUC INDICATES THE AREA UNDER THE ROC CURVE, A METRIC FOR BINARY CLASSIFICATION.

Method		AUC
Non-RNN	Zero	0.8319 (0.0070)
	Mean	0.8189 (0.0085)
	LOCF	0.8380 (0.0041)
2-stage	GRUI	0.8072 (0.0093)
	E^2GAN	0.8329 (0.0092)
End-to-end	GRU-D	0.8297 (0.0069)
	BRITS	0.8575 (0.0040)
	Ours	**0.8649 (0.0020)**

- **The mortality prediction performances of ablated models**
  - Our model outperforms ablated models, indicating that every module of the proposed model has crucial roles in performing the downstream task.
  - For example, the proposed decaying mechanism improves the model performance significantly since it helps the model consider the time gap between observations appropriately.

TABLE III
ABLATION STUDY RESULT OF OUR PROPOSED METHOD ON THE PHYSIONET DATASET.

Models	AUC
Ours	**0.8649 (0.0020)**
Ours w/o adversarial learning	0.8621 (0.0053)
Ours w/o end-to-end learning	0.8488 (0.0050)
Ours w/o dropping mechanism	0.8580 (0.0045)
Ours w/o input decaying	0.8548 (0.0079)
Ours w/o hidden vector decaying	0.8505 (0.0066)
Ours w/o gating module 1	0.8555 (0.0073)
Ours w/o gating module 2	0.8477 (0.0029)

Experimental results

- **Effectiveness of the proposed dropping mechanism on the synthetic dataset**
  - Removing and reconstructing the data improves the model performance both in terms of the missing value imputation and the downstream classification.
  - Regardless of the additional missing rate, the method significantly increases the imputation accuracy for unknown variables.
  - MSE w/o noise is smaller than MSE w/ noise, indicating that our proposed model successfully captures the true data distribution removing the noise in the data.

TABLE IV

EXPERIMENTAL RESULTS WITH VARIOUS ADDITIONAL MISSING RATE α ON THE SYNTHETIC DATASET. WE REPORT THE AUC SCORE OF THE CLASSIFICATION TASK AND THE SQUARED DISTANCE ERROR ON THE DATA WITH NOISE AND WITHOUT NOISE, NAMED MSE WITH NOISE WITH NOISE AND MSE W/O NOISE, RESPECTIVELY.

		Deterministic dropping					Dynamic dropping				
Additional missing rate	0%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
AUC	0.939 (0.005)	0.949 (0.007)	**0.950 (0.006)**	0.945 (0.010)	0.944 (0.009)	0.942 (0.008)	0.948 (0.008)	0.945 (0.006)	0.942 (0.005)	0.941 (0.005)	0.940 (0.006)
MSE w/ noise	1.106 (0.031)	1.012 (0.042)	1.021 (0.045)	1.000 (0.071)	0.988 (0.070)	0.992 (0.077)	1.046 (0.037)	1.004 (0.036)	0.994 (0.050)	0.984 (0.068)	**0.983 (0.076)**
MSE w/o noise	0.911 (0.026)	0.818 (0.041)	0.827 (0.043)	0.805 (0.071)	0.794 (0.069)	0.799 (0.076)	0.851 (0.037)	0.810 (0.034)	0.800 (0.049)	0.790 (0.067)	**0.790 (0.074)**

Thank you