A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data

SNU AI 장민석

Introduction(1/3)

- Monitoring the behaviors of these systems generates a substantial amount of multivariate time series data
  - the readings of the networked sensors (e.g., temperature and pressure) distributed in a power plant
  - the connected components (e.g., CPU usage and disk I/O) in an Information Technology (IT) system
- A critical task in managing these systems is to detect anomalies in certain time steps such that the operators can take further actions to resolve underlying issues
  - power plant failure, financial, manufacturing plant etc.

Introduction(2/3)

Previous Solution

- Distance/clustering methods
- Probabilistic methods
- Density estimation method
- Temporal prediction approaches
- Deep learning techniques

Introduction(3/3)

- Main Problem
  - Few or even no anomaly label is available in the historical data, which makes the supervised algorithms infeasible.

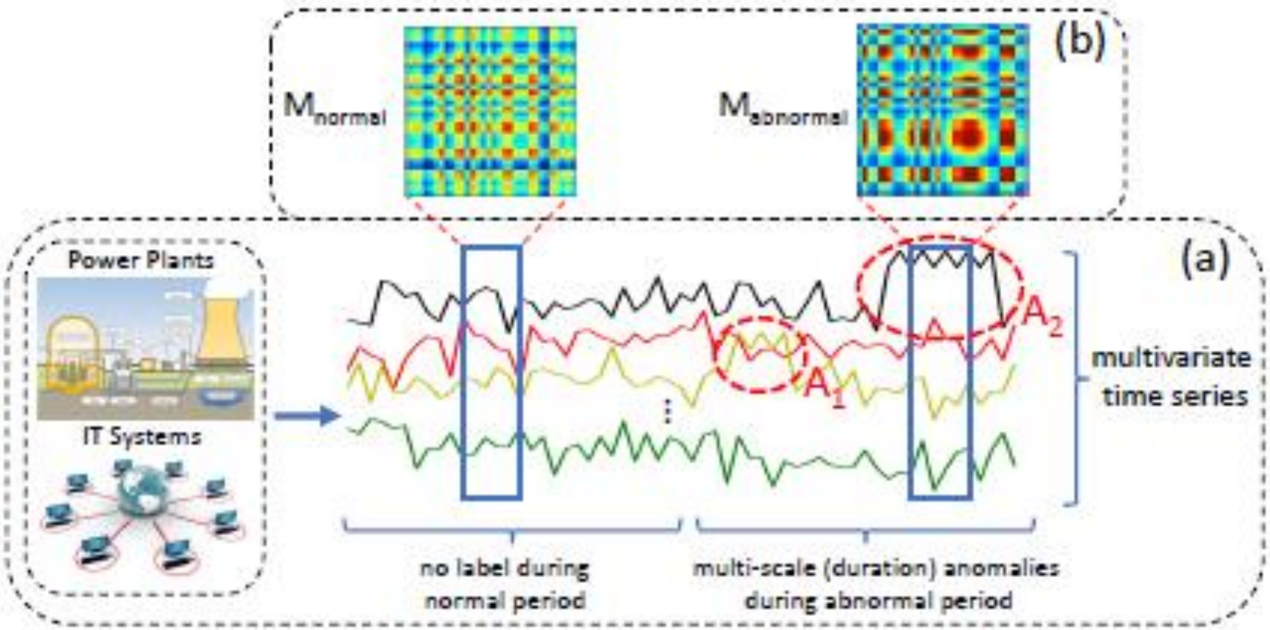
- 3 reasons that most of them may still not be able to detect anomalies effectively
  - There exists temporal dependency in multivariate time series data.
  - Multivariate time series data usually contain noise in real word applications.
  - In real world application, it is meaningful to provide operators with different levels of anomaly scores based upon the severity of different incidents.

MSCRED(1/2)

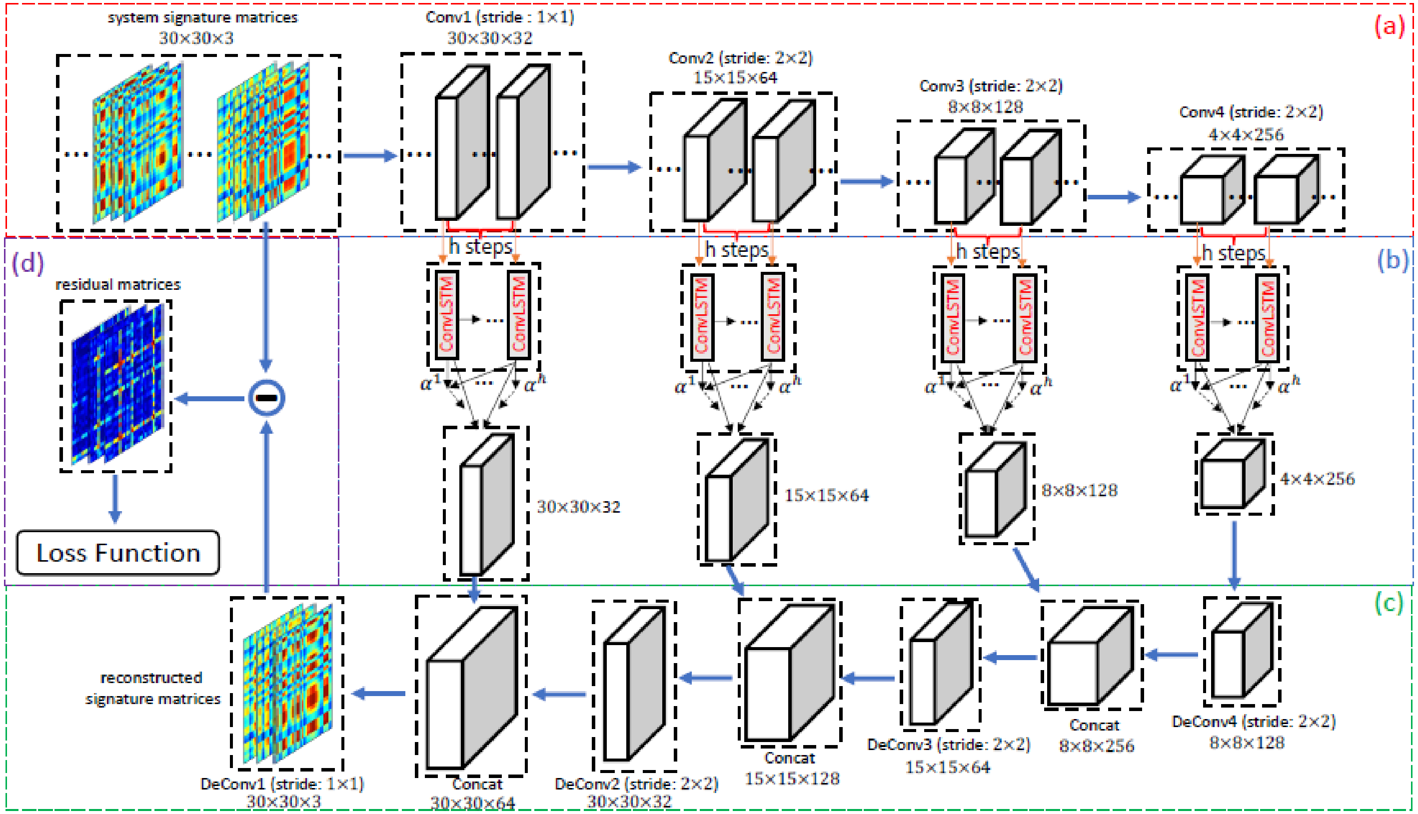
- Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED)
  - (Step1) it constructs multi-scale (resolution) signature matrices to characterize multiple levels of the system statuses across different time steps.
  - (Step2) Subsequently, given the signature matrices, a convolutional encoder is employed to encode the inter-sensor(time series) correlations patterns
  - (Step3) an attention based Convolutional Long-Short Term Memory (ConvLSTM) network is developed to capture the temporal patterns.
  - (Step4) the feature maps which encode the inter-sensor correlations and temporal information, a convolutional decoder is used to reconstruct the signature matrices
  - (Step5) the residual signature matrices are further utilized to detect and diagnose anomalies.

MSCRED(2/2)

- Core Idea of MSCRED
  - MSCRED cannot reconstruct Mabnormal well as training matrices (e.g.,Mnormal) are distinct from Mabnormal
- Included Technique
  - Fully convolutional NN, Convolutional LSTM, Attention



MSCRED Framework(1/8)



(a) signature matrices + fully convolutional

(b) Temporal patterns modeling(attention based convolutional LSTM)

(c) signature matrices decoding (deconvolutional)

(d) loss function

MSCRED Framework(2/8)

- Problem Statement

Given the historical data of n time series with length T, *i.e.*, $X = (x_1, \ldots, x_n)^T \in R^{n \times T}$ and assuming that there exists no anomaly in the data, we aim to achieve two goals:

- **Anomaly detection,** *i.e.,* detecting anomaly events at certain time steps after T.
- **Anomaly diagnosis,** *i.e.,* given the detection results, identifying the abnormal time series that are most likely to be the causes of each anomaly and interpreting the anomaly severity (duration scale) qualitatively.

MSCRED Framework(3/8)

Characterizing Status with Signature Matrices(M^t)

- the correlations between different pairs of time series are critical to characterize the system status.
- the inter-correlations between different pairs of time series in a multivariate time series segment from *t-w to t (In this paper, w = 10, 20, 30[s=3])*

$$x_i^w = (x_i^{t-w}, x_i^{t-w-1}, \dots, x_i^t)$$

$$x_j^w = (x_j^{t-w}, x_j^{t-w-1}, \dots, x_j^t)$$

X

$$m_{ij}^t = \frac{\sum_{\delta=0}^{w} x_i^{t-\delta} x_j^{t-\delta}}{\kappa}$$

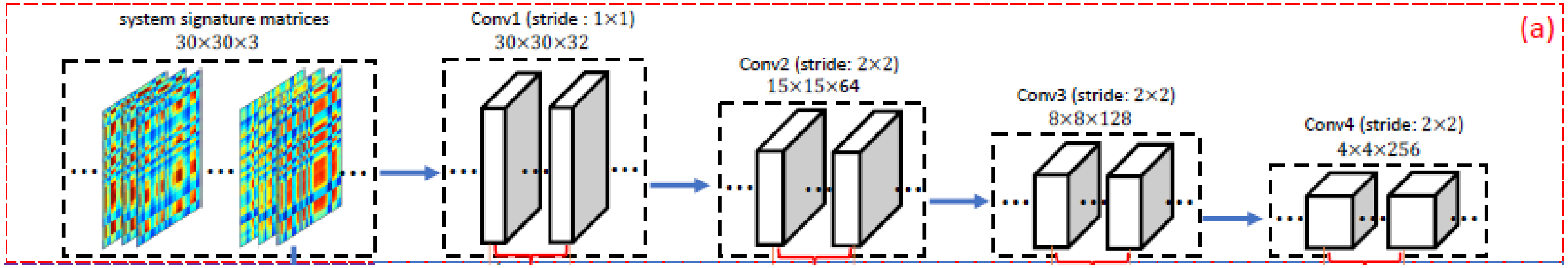
rescale factor(=window)

$m_{ij}^t \in M^t$

- Signature matrix capture the shape similarities and value scale correlations between two time series
- Signature matrix is robust to input noise as the turbulence at certain time series has little impact on the signature matrices

MSCRED Framework(4/8)

- Convolutional Encoder with CNN



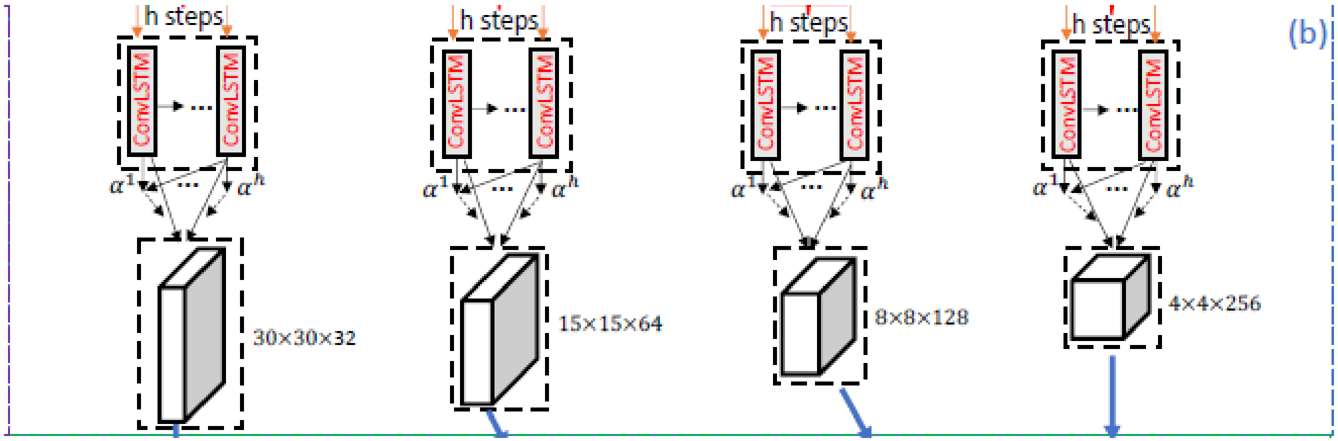
- A fully convolutional encoder is used to encode the spatial pattern of each system's signature matrices.
- the feature maps in the (l -1)th layer, the output of l-th layer

$$\mathcal{X}^{t,l} = f(W^l * \mathcal{X}^{t,l-1} + b^l)$$

- Note that the exact order of the time series based on which the signature matrices are formed is not important, because for any given permutation, the resulting local patterns can be captured by the convolutional encoder

MSCRED Framework(5/8)

- Attention based ConvLSTM(1/2)

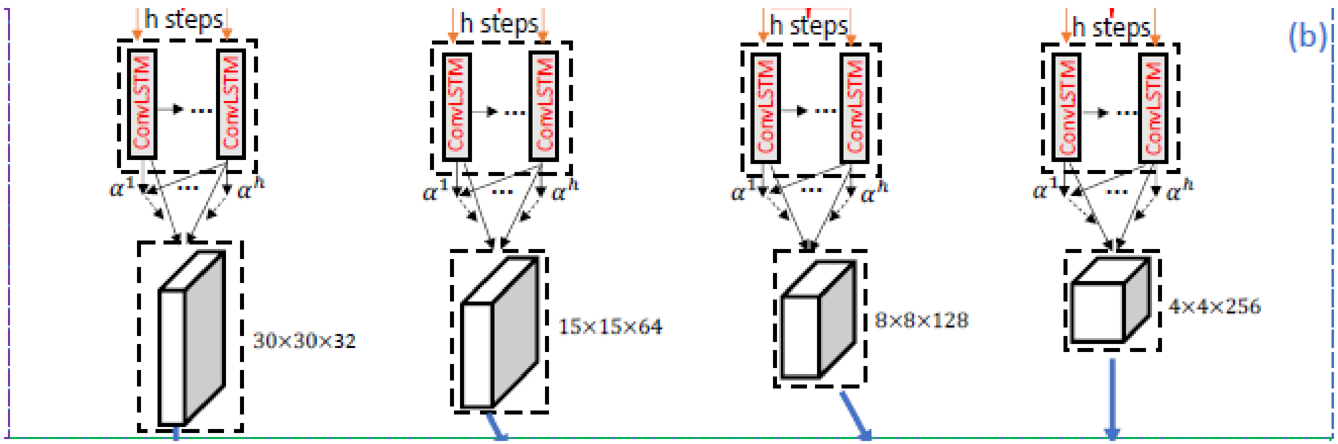


- The spatial feature maps generated by convolutional encoder is temporally dependent on previous time steps.

- ConvLSTM capture the temporal information
  - padding = (kernel size – 1) / 2 for uniforming size
  - Processes both spatial and temporal information
  - Reduce the number of weights in the model (vs LSTM)

MSCRED Framework(6/8)

- Attention based ConvLSTM(2/2)



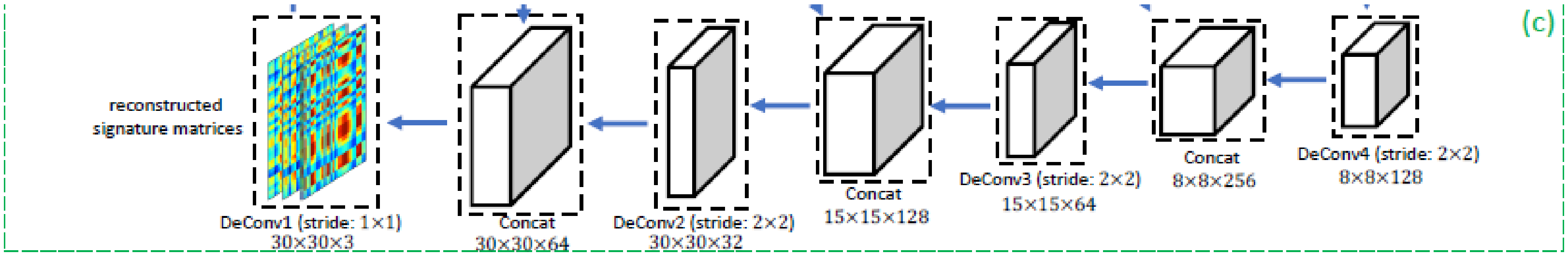
- Attention address performance deteriorate as the length of sequence increases.
- Considering not all previous steps are equally correlated to the current state($\hat{\mathcal{H}}^{t,l}$)

$$\hat{\mathcal{H}}^{t,l} = \sum_{i \in (t-h,t)} \alpha^i \mathcal{H}^{i,l}, \alpha^i = \frac{\exp\{\frac{\mathrm{Vec}(\mathcal{H}^{t,l})^{\mathrm{T}}\mathrm{Vec}(\mathcal{H}^{i,l})}{\chi}\}}{\sum_{i \in (t-h,t)} \exp\{\frac{\mathrm{Vec}(\mathcal{H}^{t,l})^{\mathrm{T}}\mathrm{Vec}(\mathcal{H}^{i,l})}{\chi}\}}$$

- Summary : the attention based ConvLSTM jointly models the spatial patterns of signature matrices with temporal information at each convolutional layer

MSCRED Framework(7/8)

- Convolutional Decoder



- To decode the feature maps obtained in previous step and get the reconstructed signature matrices

deconvolution operation

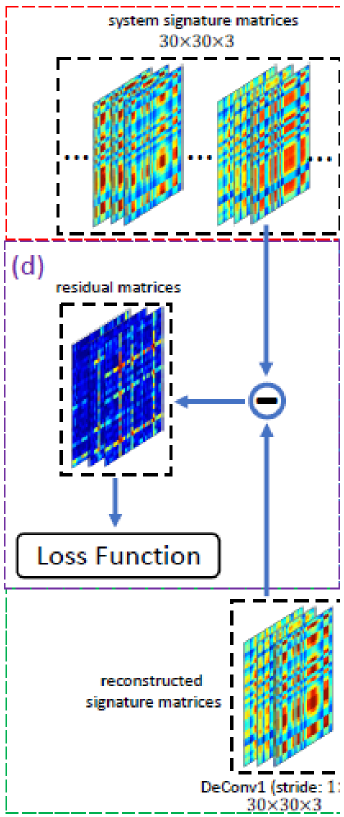
$$\hat{\mathcal{X}}^{t,l-1} = \begin{cases} f(\hat{W}^{t,l} \circledast \hat{\mathcal{H}}^{t,l} + \hat{b}^{t,l}), & l = 4 \\ f(\hat{W}^{t,l} \circledast [\hat{\mathcal{H}}^{t,l} \oplus \hat{\mathcal{X}}^{t,l}] + \hat{b}^{t,l}), & l = 3,2,1 \end{cases}$$

concatenation operation

- f() is the activation unit(same as the encoder)

MSCRED Framework(8/8)

Loss Function



- Residual matrices

$$\mathcal{L}_{MSCRED} = \sum_{t} \sum_{c=1}^{s} \left\| \mathcal{X}^{t,0}_{:,:,c} - \hat{\mathcal{X}}^{t,0}_{:,:,c} \right\|_F^2$$

- Loss function

Sum Resiual matrices of element-wise square

Experiments(1/6)

- research questions

➢ Detection
- RQ1 : Outperform baseline methods for anomaly detection in multivariate time series?
- RQ2 : How does each component of MSCRED affect its performance ?

➢ Diagnosis
- RQ3 : Whether MSCRED can perform root cause identification ?
- RQ4 : Anomaly severity (duration) interpretation effectively?

➢ Robustness to noise
- RQ5 : Compared with baseline methods, whether MSCRED is more robust to input noise ?

Experiments(2/6)

Data Setup

- Synthectic data.

$$\mathcal{S}(t) = \begin{cases} \underbrace{sin}_{C1}\underbrace{[(t-t_0)/\omega]}_{C2} + \underbrace{\lambda \cdot \epsilon}_{C3}, & s_{\text{rand}} = 0 \\ \underbrace{cos}_{C1}\underbrace{[(t-t_0)/\omega]}_{C2} + \underbrace{\lambda \cdot \epsilon}_{C3}, & s_{\text{rand}} = 1 \end{cases}$$

- C1 : Temporal pattern simulation of trigonometric
- C2 : Time delay, frequency
- C3 : random Gausian noise (scaled by factor 0.3)

- Power plant data.
- 36 time series genenrated by sensors distributed in the power plant system.
- 23,040 time steps and contains one anomaly identified by the system operator
- inject 4 additional anomalies into the test period

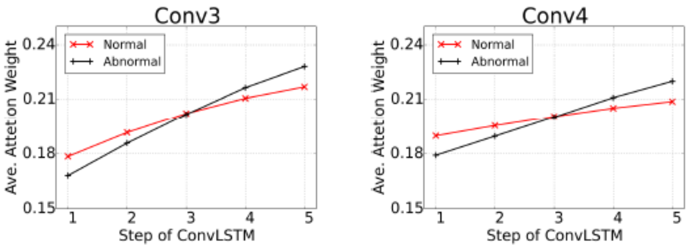
Statistics	Synthetic	Power Plant
# time series	30	36
# points	20,000	23,040
# anomalies	5	5
# root causes	3	3
train period	$0 \sim 8{,}000$	$0 \sim 10{,}080$
valid period	$8{,}001 \sim 10{,}000$	$10{,}081 \sim 18{,}720$
test period	$10{,}001 \sim 20{,}000$	$18{,}721 \sim 23{,}040$

Experiments(3/6)

- Result of R1, R2
  - Detection
  - RQ1 : Outperform baseline methods for anomaly detection in multivariate time series?
  - RQ2 : How does each component of MSCRED affect its performance ?

Method	Synthetic Data			Power Plant Data		
	Pre	Rec	F_1	Pre	Rec	F_1
OC-SVM	0.14	0.44	0.22	0.11	0.28	0.16
DAGMM	0.33	0.20	0.25	0.26	0.20	0.23
HA	0.71	0.52	0.60	0.48	0.52	0.50
ARMA	0.91	0.52	0.66	0.58	0.60	0.59
LSTM-ED	1.00	0.56	0.72	0.75	0.68	0.71
ⓐ $\mathrm{CNN}^{ED(4)}_{ConvLSTM}$	0.37	0.24	0.29	0.67	0.56	0.61
ⓑ $\mathrm{CNN}^{ED(3,4)}_{ConvLSTM}$	0.63	0.56	0.59	0.80	0.72	0.76
ⓒ $\mathrm{CNN}^{ED}_{ConvLSTM}$	0.80	0.76	0.78	0.85	0.72	0.78
MSCRED	**1.00**	**0.80**	**0.89**	**0.85**	**0.80**	**0.82**
Gain (%)	–	30.0	23.8	13.3	19.4	15.5

ⓐ without attention
ⓑ without attention and two ConvLSTM Layer
ⓑ without attention and three ConvLSTM Layer



the average distribution of attention weghts over 5 previous timesteps at last two ConvLSTM Layers

- the latter distribution, the older timesteps (step 1or 2), which tend to still be normal and therefor in a different system status than current timestep(step 5), are assigned lower weights than in the distribution for normal segments.
- the attention modules show high sensitivity

Experiments(4/6)

- Result of R3,R4(1/2)
  - ➢ Diagnosis
  - ● RQ3 : Whether MSCRED can perform root cause identification ?

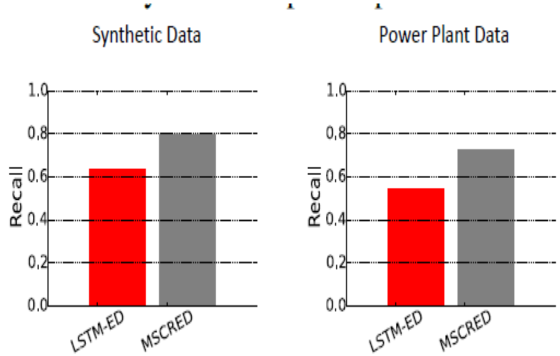


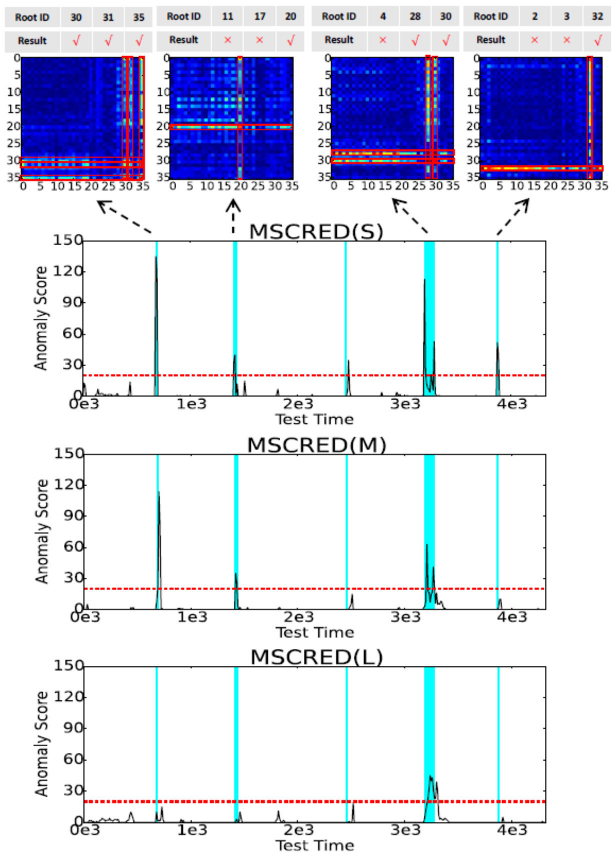
Figure 5: Performance of root cause identification.

- ● Root cause ? cause of Alarm
- ● Performance better than best baseline LSTM-ED

Experiments(5/6)

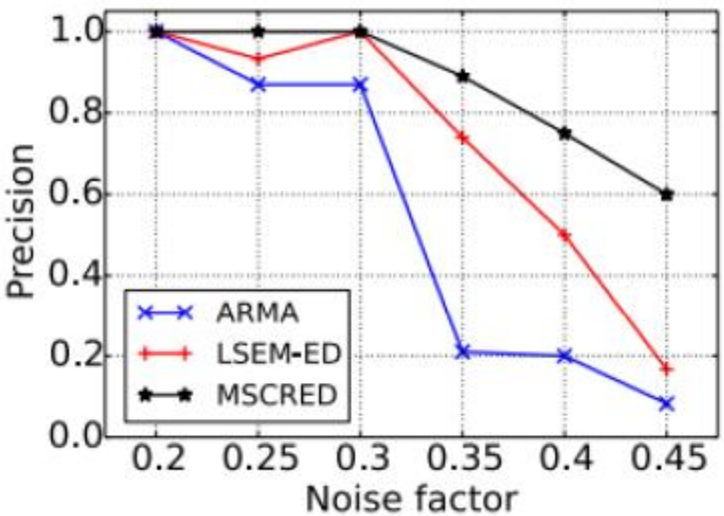
- Result of R3,R4(1/2)
  - ➢ Diagnosis
  - RQ4 : Anomaly severity (duration) interpretation effectively?

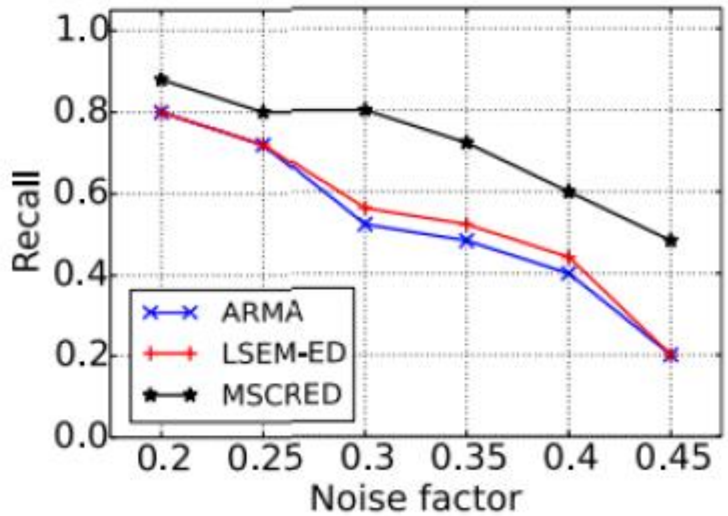
- If the segment size is small, it is possible to detect all anomaly from short duration to long anomaly.
- As the segment size increases, only long duration anomaly can be detected.
- Severity is defined as duration, and can be interpreted according to severity beyond anomaly



Experiments(5/6)

- Result of R5
  - Robustness to noise
  - RQ5 : Compared with baseline methods, whether MSCRED is more robust to input noise ?





- MSCRED has high precision and recall values both when the noise scale is small and when the noise scale is large.

T h a n k Y o u