Supplementary Materials: Deep Multi-Instance Contrastive Learning with Dual Attention for Anomaly Precursor Detection

Dongkuan Xu*† Wei Cheng‡ Jingchao Ni‡ Dongsheng Luo* Masanao Natsumeda‡
Dongjin Song§ Bo Zong‡ Haifeng Chen‡ Xiang Zhang*

1 Notations

Table 1: Notations

Notation	Meaning
N	Number of sensory variables
d	Hidden feature dimension for each sensory variable
x_t^l	Value of the l-th sensory variable at time step t
$\boldsymbol{\alpha} \in \mathbb{R}^n$	Attention values for different instances
$\boldsymbol{\beta}_k \in \mathbb{R}^N$	Attention values for different variables
$\mathbf{x}_t \in \mathbb{R}^N$	A vector of N time series at time t
$\mathbf{X} \in \mathbb{R}^{N \times T}$	Multivariate time series
$\mathbf{E} \in \mathbb{R}^{N \times I}$	An instance / A time series segment
$\mathbb{B} = \{\mathbf{E}_1, \cdots\}$	A bag / A set of instances
$\mathbf{G} \in \mathbb{R}^{N \times d}$	Transformed representation of an instance
$\mathbf{Q} \in \mathbb{R}^{N \times d}$	Transformed representation of a bag
$\mathbf{Z} \in \mathbb{R}^{M \times I}$	Anomaly precursor

The main notations are summarized in Table 1.

2 Computational Analysis

LSTM is local in time [1] and its time complexity per parameter is $\mathcal{O}(1)$ for each time step. Thus the overall complexity of LSTM per time step is proportional to the number of parameters. The parameters of MCDA are from the LSTM unit and the dual attention mechanism. Specifically, the parameters of LSTM are from the calculation of the cell updating matrix, the forget gate, the input gate and the output gate. So the overall complexity for our LSTM unit per time step is $\mathcal{O}(N^2d^2 + N^3d)$. The parameter numbers for the instance attention and the variable attention are $S + 2SNd$ and $\tilde{S} + 2\tilde{S}d$ respectively. Thus, the overall complexity per time step of the dual attention is $\mathcal{O}(SNd)$. Because $S = \tilde{S}$, $N > d$ and $N > S$, the overall complexity per time step of MCDA is $\mathcal{O}(N^3)$. Note that the majority of parameters are from the gate calculation and are related to the correlation term $\mathbf{M}_t$. Usually most of the variable correlations are not related to the anomaly and we can focus on a small part of them. Thus we can apply the tricks like parameter sparsity [2] and parameter factorization [3] to reduce the number of parameters into $\mathcal{O}(N^2)$ practically.

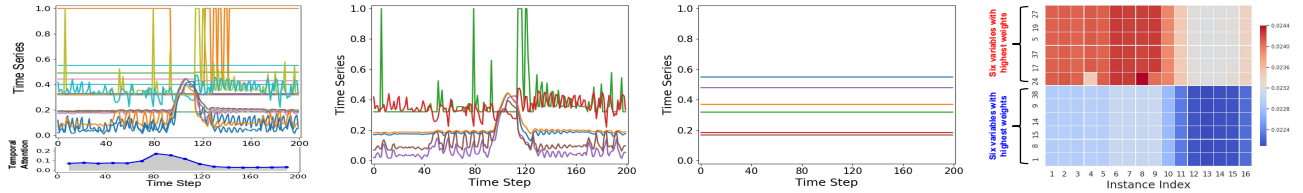
3 Additional Experimental Results

3.1 Precursor Detection on Another Positive Bag from Showcase data The upper part of Fig. 1(a) is the original time series, the bottom part is the attention weights inferred. The precursor is located in the middle several time steps. Similar observations can be observed in Fig. 1 as in Fig. 6 in the main paper.

3.2 Task 3 on Synthetic Data To gain further insight of our method, we construct the synthetic data with the anomaly precursor that is resulted from the change of correlations between different sensory variables. The synthetic data is shown in the upper part of Fig. 2(a) or 2(c). It contains six sensory variables. From top to bottom, the first and third variables are almost constant, and the second and forth ones are periodic. The bottom two are random but share the similar temporal patterns. The precursor lies in the first several time steps and is involved in the bottom two variables, because the bottom two variables becomes independent at the beginning. From Fig. 2(a), it is observed that the attention values of the period of precursor is much larger, which demonstrates MCDA detects the time location of the precursor successfully. The variable attention values of the six variables across different instances are shown in Fig. 2(b). It is observed that the variable attention values of the bottom two are larger than the ones of others, especially in the first two instances. This validates the ability of MCDA to detect the precursor of the anomaly resulted from the change of correlations between variables.

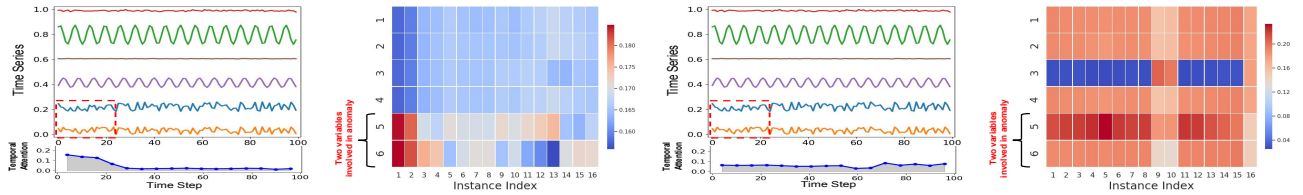
3.3 Results on Cyber-Physical System Data Another real data we used is a real cyber-physical system data [4] generated from manufacturing industry. This dataset contains time series collected from 1625 electric sensors installed on different components of the

*Penn State University. {dux19, dul262, xzz89}@psu.edu
†Work done during an internship at NEC Labs America
‡NEC Labs America, Inc. {weicheng, jni, mnatsumeda, bzong, haifeng}@nec-labs.com
§University of Connecticut. {dongjin.song}@uconn.edu



(a) Temporal attention distribution across time steps (b) Six time series (variables) with highest attention values (c) Six time series (variables) with lowest attention values (d) Variable attention distribution across instances

Figure 1: The precursor detection result on another positive bag from the Showcase data. The precursor is located in the middle several time steps.



(a) Temporal attention generated by MCDA (b) Variable attention generated by MCDA (c) Temporal attention generated by MCDA-M (d) Variable attention generated by MCDA-M

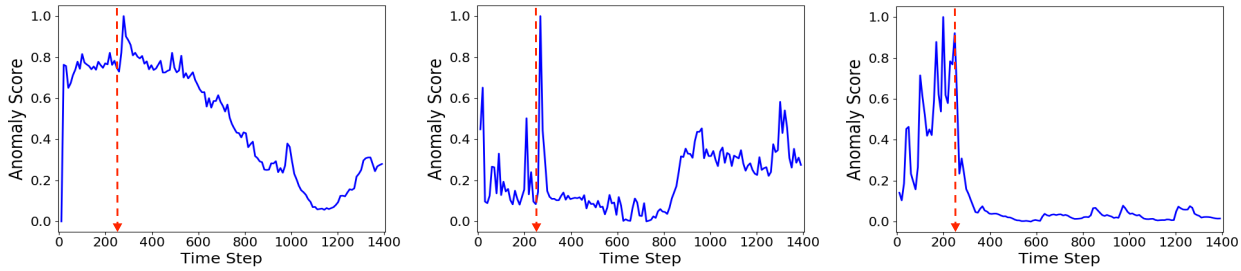
Figure 2: MCDA vs MCDA-M on the synthetic data. The precursor resulted from the change of correlations is located in the first several time steps.

Cyber-physical system. The total length is 1400 time steps. The anomaly occurred at the 210-th time step. Fig. 3 show the results on the Cyber-physical system data. The red dots indicate the time of anomalies labeled by users. Higher anomaly score indicates higher probability of anomaly. We use the one reported anomaly as the training data for MCDA and show the detection results on the whole time length data. As can be seen in Figs. 3(a) and 3(b), MCDA can detect the anomaly accurately. LSTM-AE and DAGMM mistakenly regard the time periods after the 210-th and the 1000-th time step as the anomaly periods. Moreover, it is noted that the detected anomaly period of MCDA is a little earlier than the 210-th time step, which verifies the effectiveness of MCDA to detect the incipient faults of the anomaly.

3.4 Visualization of Synthetic Datasets In order to evaluate the ability of MCDA to detect different kinds of precursors, we construct six different synthetic datasets as shown in Fig. 4. The time length of each dataset is 1500 time steps. Each dataset contains two annotated precursors and the anomaly is assumed to happen after precursors. The two precursors in the six datasets cover the same time period. For the first four datasets, there are five sensory variables. The precursor is assumed to occur because of the pattern change of the last variable (from top to bottom). The last two datasets contains six sensory variables. Synthetic Data 5 is the dataset described in the main paper.

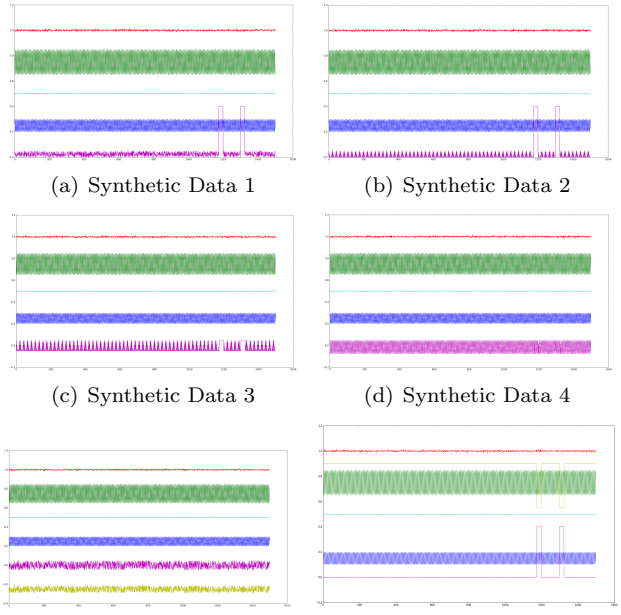
Let's take Synthetic Data 6 as an example to describe how to generate these synthetic datasets. Synthetic Data 6 contains 6 variables and the total time steps are 1500 as shown in Fig. 5. Two variables are constant signals (1.0 and 0.5) added with different noises (Normal(0, 0.005), Normal(0, 0.001)). Two variables are generated from sine and cosine functions. Their value ranges are rescaled as [0.1, 0.2] and [0.65, 0.85]. Another two variables are the target variables. Both of them have some significant changes at two same time steps. Their values are constants (0.0 and 0.9) added with the same noise (Normal(0, 001)). Two positive bags are created as shown in Fig. 5(a)-5(b). Positive Bag 1 contains a precursor located in the last several time steps and Positive Bag 2 contains a precursor located in the first several time steps. The two target variables show the same change at the two time steps respectively, which indicates that the two anomaly precursors are the same.

3.5 Interpreting Precursors on Synthetic Data 6 We show the detection results of MCDA on Synthetic Data 6. Two detected most related variables are shown in Fig. 5(c)-5(d). The attention values of the 16 instances for Positive Bag 1 are [0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0624181, 0.0628514, 0.0628557, 0.0628564]. The at-



(a) LSTM-AE (b) DAGMM (c) MCDA

Figure 3: Anomaly detection results on the Cyber-physical System data. Red dots indicate the time of anomalies labeled by users.



(a) Synthetic Data 1 (b) Synthetic Data 2

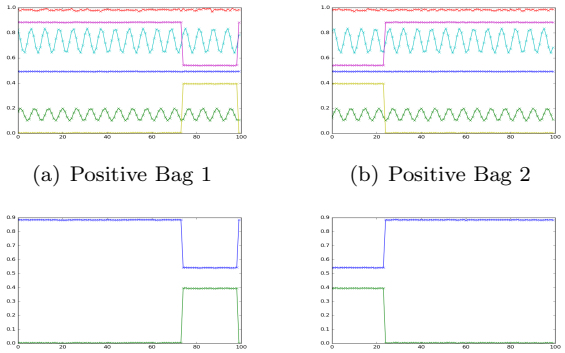
(c) Synthetic Data 3 (d) Synthetic Data 4

(e) Synthetic Data 5 (f) Synthetic Data 6

Figure 4: Synthetic Datasets

tention values of the last three instances are larger. The attention values of the 16 instances for Positive Bag 2 are [0.0629492, 0.0629477, 0.0629467, 0.0629442, 0.0629377, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976, 0.0622976]. The attention values of the fist five instances are larger. The results of attention values at instance level verifies the ability of MCDA to detect the time location of precursor, i.e., 'when'.

The variable attention values in the last instance of Positive Bag 1 are [9.58e-05, 9.95e-01, 4.40e-04, 3.74e-04, 7.13e-05, 3.69e-03]. The second and the last variables have the larger values. The variable attention values in the first instance of Positive Bag 2 are [9.59e-05, 9.95e-01, 4.39e-04, 3.74e-04, 7.12e-05, 3.70e-03]. The second and the last variables have the larger values. The results of attention values at variable level verifies the ability of MCDA to detect the involved sensory variables of precursor, i.e., 'where'.



(a) Positive Bag 1 (b) Positive Bag 2

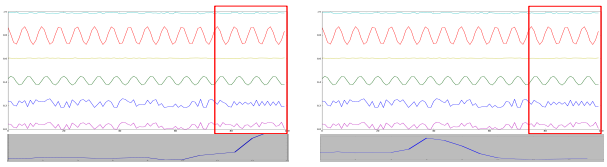
(c) Two related variables (d) Two related variables

Figure 5: Two positive bags of Synthetic Data 6.

4 Effectiveness of $\mathbf{M}_t$

To verify the effectiveness of $\mathbf{M}_t$, we add the precursor detection result on Positive Bag 2 of Synthetic Data 5. The detection result of Positive Bag 1 is shown in the main paper (Fig. 2(a)-2(b)). Positive Bag 2 is shown in the upper part of Fig. 6(a) or Fig. 6(b). From top to bottom, the first and third variables are almost constant, and the second and forth ones are periodic. The bottom two are random. They are correlated with each other at beginning but become independent later. The precursor period lies in the last several time steps that is tagged by the red rectangle in g. 6(a) and Fig. 6(b). The blue line at bottom of the figures indicates the attention values of different instances. As can be seen in Fig. 6(a), the attention values of the last several time steps are larger, which indicates MCDA detects the time location of precursors successfully. The attention values of the six variables are [0.07, 0.03, 0.03, 0.07, 0.38, 0.41], which means that MCDA regards the bottom two to be more correlated with the anomaly

event. This indicates that MCDA detects the sensor location of precursors successfully. To sum up, the ability of MCDA to detect the precursor of the anomaly resulting from the change of correlations between time series is verified. Moreover, MCDA-M did not detect the precursor successfully as shown in Fig. 6(b), which also verifies the effectiveness of the $\mathbf{M}_t$ term in MCDA.



(a) By MCDA (b) By MCDA-M

Figure 6: MCDA vs MCDA-M on Positive Bag 2 of Synthetic Data 5.

We add the ROC curves of MCDA and MCDA-M based on Showcase 28, 34 to verify the effectiveness of $\mathbf{M}_t$. Specially, we use Showcase 34 as the training set and test on the two precursor cases from Showcase 28. The results are shown in Fig. 7(a)-7(a). As can be seen in Fig. 7(a) and Fig. 7(b), MCDA outperforms MCDA-M on the task of precursor detection, which verifies the effectiveness of $\mathbf{M}_t$.

5 Parameter Sensitivity

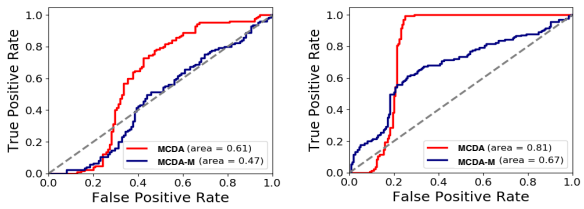
We also study the influence of λ and η on the anomaly detection. Fig. 8 shows the results on the Showcase 28 data. It is observed that MCDA detected the two anomalies successfully, which indicated that the performance of MCDA on anomaly detection is also not sensitive to λ and η.

6 Comparison Among Different Baselines

We categorize the baseline methods according to different tasks as shown in Table 2.

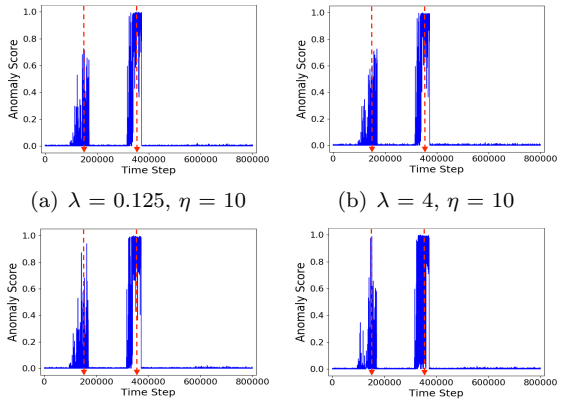
7 Discussion

7.1 The Case Where No Precursor Exists We study the anomaly precursor detection to provide the predictive maintenance for the real-world systems. The case where no precursor exists does not meet the assumption of our work. But we can use a *pre-test* strategy to determine if no precursor included in the training data. If no precursor, we can expand the time period of training set to include more data, then train the model again. We can repeat the process until precursors included. Otherwise, we will not apply the model to the test data. Specifically, to determine if no precursor included, we can apply our well trained model to the validation data, which is drawn from the same distribution of the training data. If the fitness score (see Eq. (1)) of the validation data is small, it indicates no precursor included.



(a) ROC based on SC28C1 (b) ROC based on SC28C2

Figure 7: The comparison of precursor detection performance between MCDA and MCDA-M.



(a) $\lambda = 0.125$, $\eta = 10$ (b) $\lambda = 4$, $\eta = 10$

(c) $\lambda = 0.5$, $\eta = 1$ (d) $\lambda = 0.5$, $\eta = 20$

Figure 8: Parameter sensitive analysis on anomaly detection.

7.2 The Case Where Annotated Data Is Few When developing anomaly detection methods for the real-world systems, one of the major hurdles is the lack of annotated anomalies. To handle this issue, we use the contrastive loss in the objective function. The contrast loss makes the representations of two items from the same class be similar and the ones from different classes be dissimilar. Though the annotated anomalies are few, the data from the system normal period is a lot. We sample much data from the system normal period, and use it with the annotated anomalies to construct many data pairs for the contrastive loss. If the sampled data is much enough to represent the data distribution of the system normal period, our model can effectively learn the discriminative patterns of the precursors/anomalies.

7.3 Predict When The Anomaly Will Occur Our model can help the operators conduct the predictive maintenance to avoid system anomalies. Usually it is hard to accurately predict when the next anomaly will occur. But, after trained well, our model can be used

Table 2: Comparison of baseline methods

Task	MCDA	DAGMM	LSTM-AE	Kmeans	GAKK	DTW	L2	SAX	SVM	MI-SVM
T1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
T2	✓	✓	✓							
T3	✓									

to detect the precursors in future data, then we can predict how long the anomaly will show up after the detected precursors and what the probability of the anomaly is. Specifically, we can estimate the length of the anomaly incipient period based on the labeled training data. Then based on the estimation and our detected precursors, we can tell when the anomaly will occur. In addition, the probability of the anomaly can be estimated based on the fitness score.

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] S. Narang, E. Elsen, G. Diamos, and S. Sengupta, "Exploring sparsity in recurrent neural networks," in *ICLR*, 2017.

[3] O. Kuchaiev and B. Ginsburg, "Factorization tricks for lstm networks," in *ICLR Workshop*, 2017.

[4] W. Cheng, K. Zhang, H. Chen, G. Jiang, Z. Chen, and W. Wang, "Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations," in *SIGKDD*. ACM, 2016, pp. 805–814.