Anomaly Transformer : Time series anomaly detection with association discrepancy

한양대학교 산업공학과
IDSL 석사과정 고경준

HANYANG UNIVERSITY

Contents

1. Introduction

Time-series anomaly detection

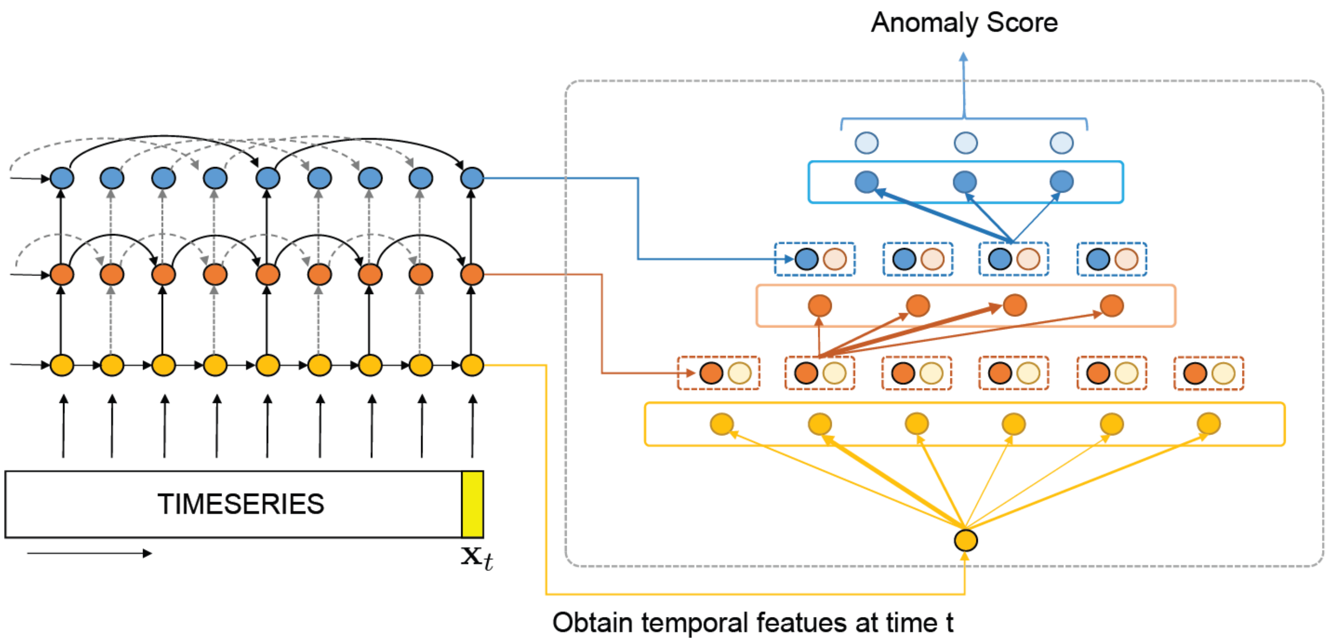
- Density-estimation methods : LOF etc.
- Clustering-based methods : OCSVM, SVDD, deep-SVDD, THOC etc.
- Reconstruction-based methods : LSTM-VAE, GAN, InterFusion etc.

Difficulties

- Various time series AD models take point-wise RNN based ways to consider the temporal information.
- But RNN based models can't provide comprehensive description of temporal context
- And point-wise representation is less informative for complex temporal patterns and can be dominated by normal time points

1. Introduction

Clustering based timeseries AD models (THOC)



- 바로 이전 시점의 정보를 반영하지 않고, skip length(s^l) 시점 이전의 정보를 반영하는 skip connection 개념을 도입
- Layer의 층 수(l)에 따라 skip length를 지수적으로 변경($s^l = 2^{l-1}$)
- 이러한 Layer를 여러 겹 쌓아서 short / long term dependency를 반영

1. Introduction

Reconstruction based timeseries AD models (InterFusion)

- **InterFusion 읽고 update.**

1. Introduction

Association discrepancy idea

- The Rarity of anomalies and the dominance of normal patterns cause difficulties to point-wise methods.
- The association of anomalies with adjacent time points is bigger than whole series.
- Prior-association : association with adjacent time points.
- Series-association : association with whole series.
- Association discrepancy : distance between prior and series associations.

Contributions covering three real applications.

- Association discrepancy를 고려할 수 있는 anomaly-attention mechanism과 anomaly-transformer를 제안.
- Minimax strategy를 반영하여 normal-abnormal distinguishability를 강화.
- Anomaly transformer가 three real applications에서 SOTA를 달성

2. Related work

2.1 Unsupervised time series anomaly detection

- Density-estimation methods : LOF, COF, DAGMM etc.
- Clustering-based methods : OCSVM, SVDD, deep-SVDD, THOC etc.
- Reconstruction-based methods : LSTM-VAE, OmniAnomaly, GANs, InterFusion etc.
- Autoregression-based methods : ARIMA, LSTMs

2.2 Transformers for time series analysis

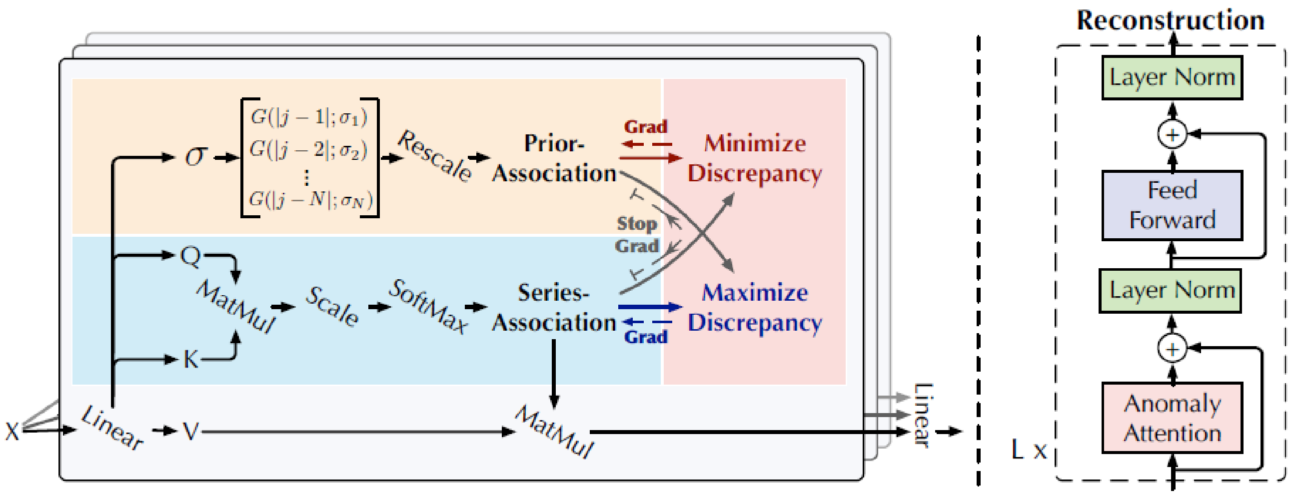
- Transformer shown great power in sequential data (NLP, audio rocessing, vision, time-series etc.)
- GTA (graph structure is adopted as well as transformer)
- Anomaly transformer renovates the self-attention mechanism

3. Method

3.0 Problem statement

- X = $\{x_1, x_2, \ldots, x_N\}$ where $x_t \in R^d$. Whether x_t is abnormal.
- Anomaly Transformer using Association discrepancy is proposed.

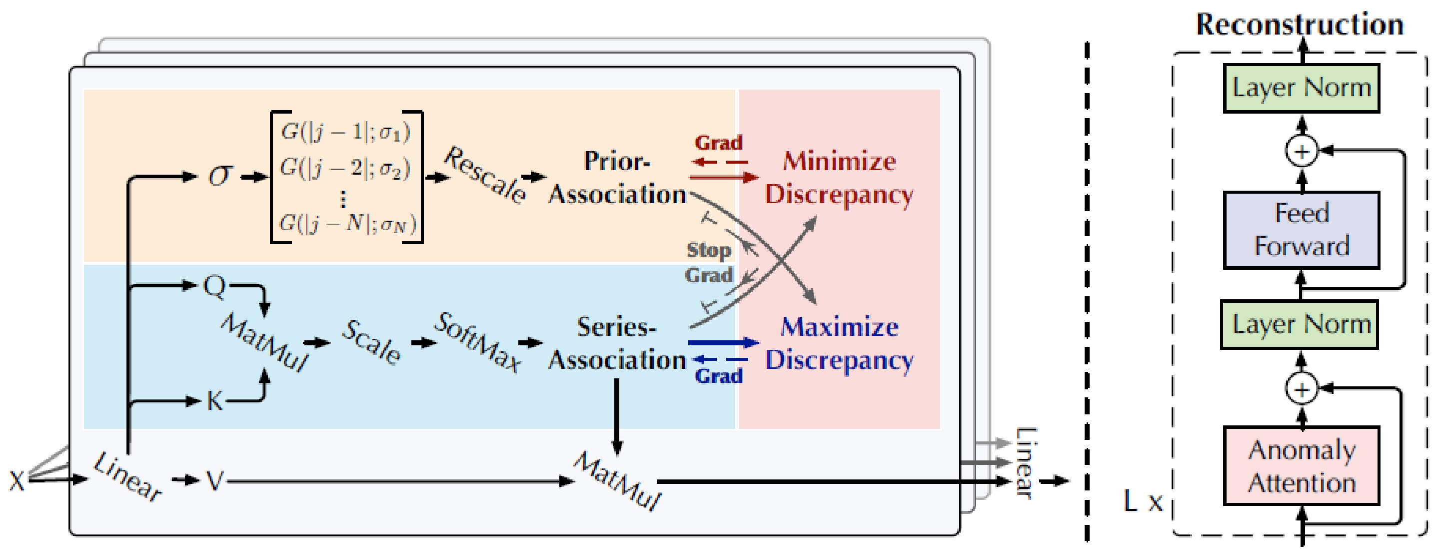
3.1 Anomaly Transformer (overall architecture)



- Stacked Anomaly Attention & Feed Forward layers with residual connection.
- In addition to the reconstruction loss, model is optimized by minimax strategy.

3. Method

3.1.1 Overall architecture (Anomaly Transformer)



$$
\begin{aligned}
\mathcal{Z}^l &= \text{Layer-Norm}\Big(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\Big) \\
\mathcal{X}^l &= \text{Layer-Norm}\Big(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\Big),
\end{aligned}
\tag{1}
$$

- Input time series is $\mathcal{X} \in \mathbb{R}^{N \times d}$,
- $\mathcal{X}^l \in \mathbb{R}^{N \times d_{\text{model}}}, l \in \{1, \cdots, L\}$, $\mathcal{X}^0 = \text{Embedding}(\mathcal{X})$.
- $d_{model} = d/\mathrm{L}$

3. Method

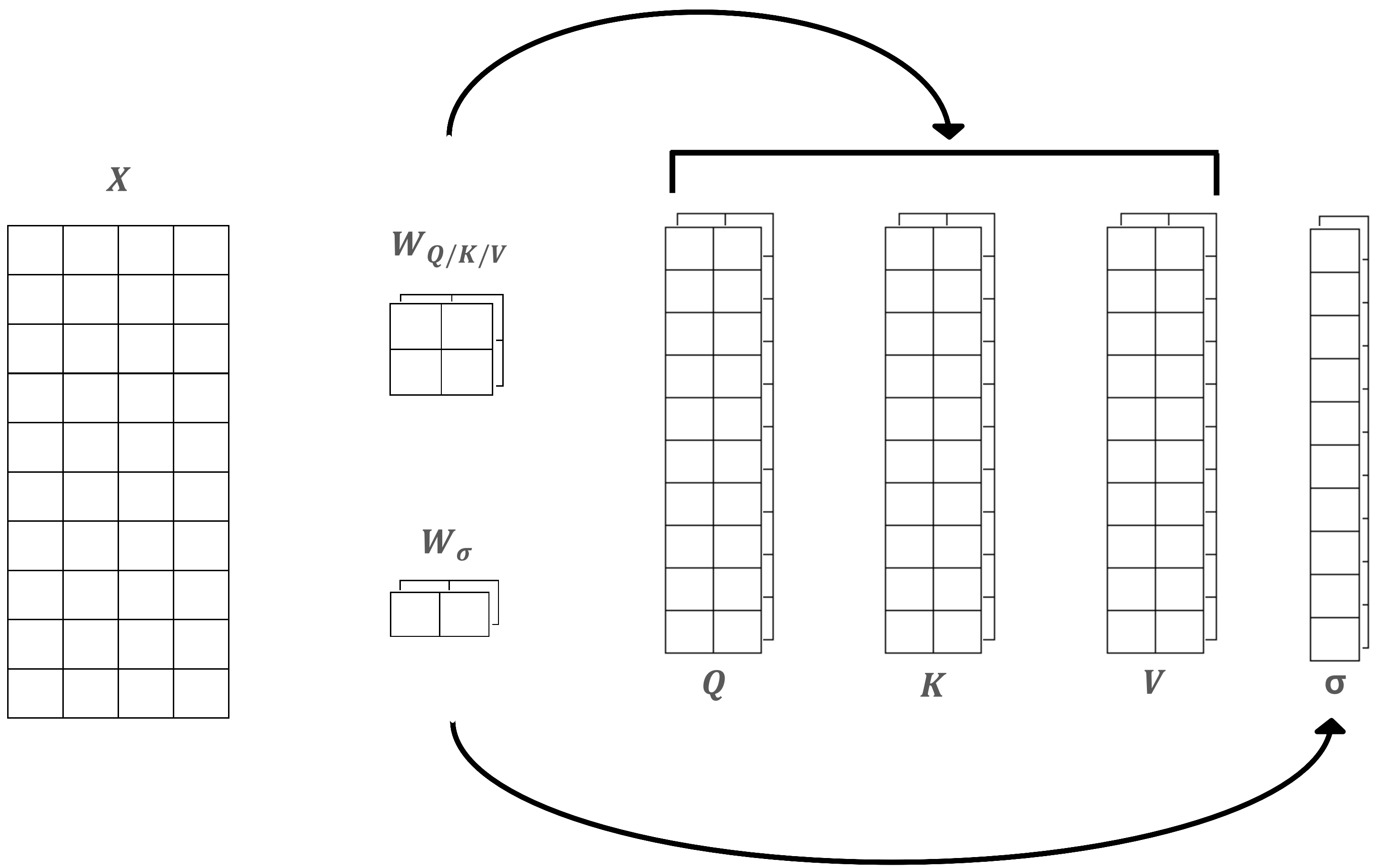
3.1.1 Anomaly attention (Anomaly Transformer)

$$
\begin{aligned}
\text{Initialization: } & \mathcal{Q}, \mathcal{K}, \mathcal{V}, \sigma = \mathcal{X}^{l-1}W_{\mathcal{Q}}^{l}, \mathcal{X}^{l-1}W_{\mathcal{K}}^{l}, \mathcal{X}^{l-1}W_{\mathcal{V}}^{l}, \mathcal{X}^{l-1}W_{\sigma}^{l} \\
\text{Prior-Association: } & \mathcal{P}^{l} = \mathrm{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{|j-i|^2}{2\sigma_i^2} \right) \right]_{i,j\in\{1,\cdots,N\}} \right) \\
\text{Series-Association: } & \mathcal{S}^{l} = \mathrm{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{model}}}} \right) \\
\text{Reconstruction: } & \widehat{\mathcal{Z}}^{l} = \mathcal{S}^{l}\mathcal{V},
\end{aligned}
\tag{2}
$$

- l-th Query, Key, Value, sigma are calculated with X^{l-1}.
- Prior-Association is calculated with learnable Gaussian kernel.
- Series-Association is calculated with outer product of query and key matrix.
- Reconstruction is calculated with series-association and value matrix.
- $d_{model} = \frac{d}{L}$, so output dimension of Anomaly-attention is equal to input dimension d

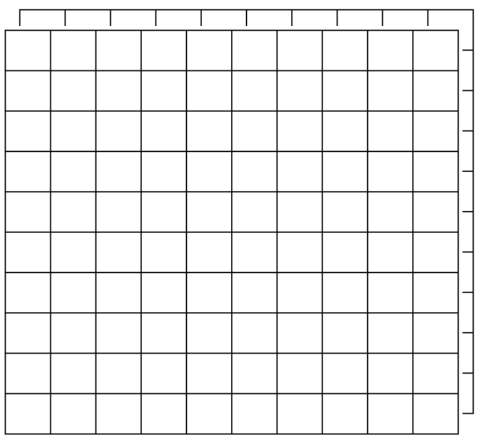
3. Method

3.1.1 Anomaly attention (Anomaly Transformer)

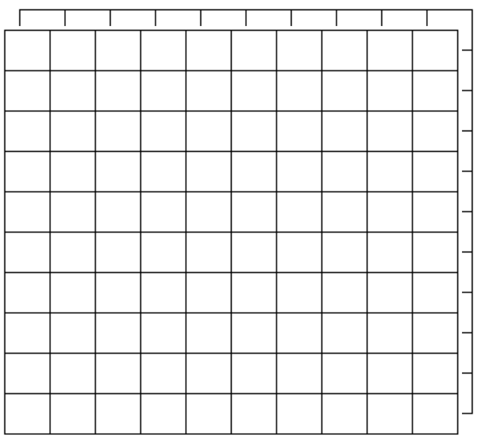


3. Method

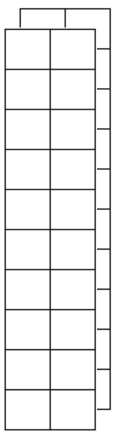
3.1.1 Anomaly attention (Anomaly Transformer)



$$S^l = Softmax(\frac{QK^T}{\sqrt{d_{model}}})$$



$$P^l$$



$$\widehat{Z}^l = S^l V^l$$

3. Method

3.1.2 Association discrepancy (Anomaly Transformer)

$$\mathrm{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X}) = \left[\frac{1}{L}\sum_{l=1}^{L}\left(\mathrm{KL}(\mathcal{P}_{i,:}^{l}\|\mathcal{S}_{i,:}^{l}) + \mathrm{KL}(\mathcal{S}_{i,:}^{l}\|\mathcal{P}_{i,:}^{l})\right)\right]_{i=1,\cdots,N} \qquad (3)$$

- **Association discrepancy between prior-series association from multiple layers**
- $\boldsymbol{AssDis(P,S;X)\ \in\ R^{N\times 1}}$
- **Anomalies will present smaller AssDis than normal time points**

$$\mathcal{L}_{\mathrm{Total}}(\widehat{\mathcal{X}},\mathcal{P},\mathcal{S},\lambda;\mathcal{X}) = \|\mathcal{X}-\widehat{\mathcal{X}}\|_{\mathrm{F}}^{2} - \lambda \times \|\mathrm{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\|_{1} \qquad (4)$$

- **Total loss function(4) contains reconstruction error and Association discrepancy**

3. Method

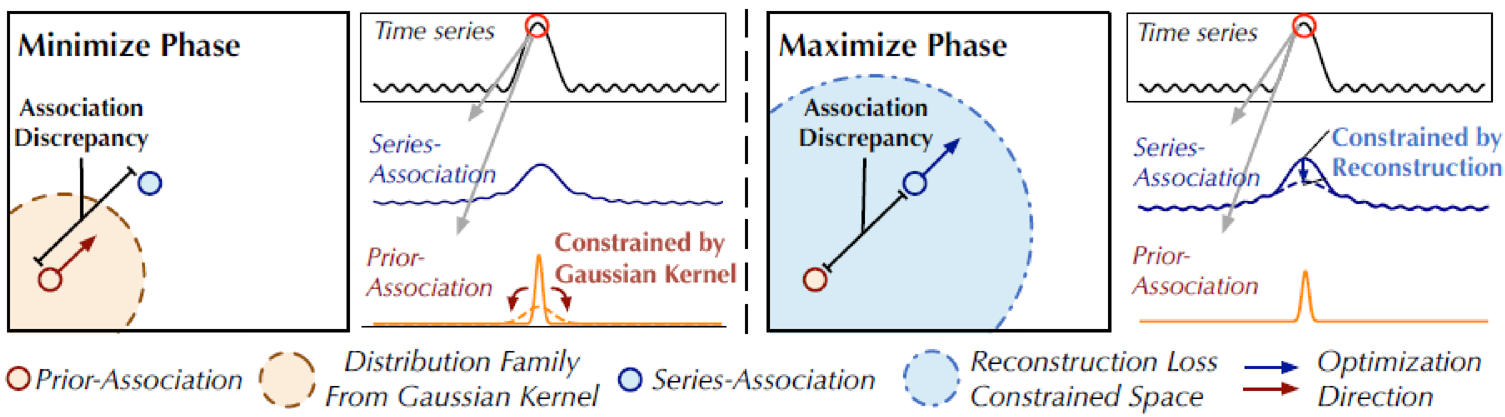
3.2.1 Minimax strategy (Minimax association learning)

$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}, \lambda; \mathcal{X}) = \|\mathcal{X} - \widehat{\mathcal{X}}\|_{\text{F}}^2 - \lambda \times \|\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\|_1 \quad (4)$$

- 제약 없이 AssDis를 maximize하면 scale parameter of Gaussian (σ)가 과도하게 작아짐.
- 이를 해결하기 위해 Minimax strategy 도입.

(series를 고정하고 prior 학습 and prior 고정하고 series 학습)

$$\begin{aligned} &\text{Minimize Phase: } \mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}_{\text{detach}}, -\lambda; \mathcal{X}) \\ &\text{Maximize Phase: } \mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}_{\text{detach}}, \mathcal{S}, \lambda; \mathcal{X}), \end{aligned} \quad (5)$$



3.2.2 Association-based anomaly criterion (Minimax association learning)

$$\text{AnomalyScore}(\mathcal{X}) = \text{Softmax}\Big(-\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})\Big) \odot \Big[\|\mathcal{X}_{i,:} - \widehat{\mathcal{X}}_{i,:}\|_2^2\Big]_{i=1,\cdots,N} \qquad (6)$$

- **Softmax of AssDis와 reconstruction error의 element-wise sum을 이용하여 Anomaly Score 도출**
- **Window 안의 abnormal point가 많을수록 AssDis와 reconstruction error가 늘어나서 큰 Anomaly Score가 산출됨**

4.0 Setting

- **Datasets**

  SMD, PSM, Both MSL, SWaT, NeurIPS-TS

- **Implementation details**

  Non-overlapped sliding window, window size : 100.

- **Baselines**

  Reconstruction based : InterFusion, BeatGAN, OmniAnomaly, LSTM-VAE

  Clustering based : ITAD, THOC, Deep-SVDD

  Autoregression based : CL-MPPCA, LSTM, VAR

  Classic methods : OC-SVM, IsolationForest

4. Experiments

4.1 Main results

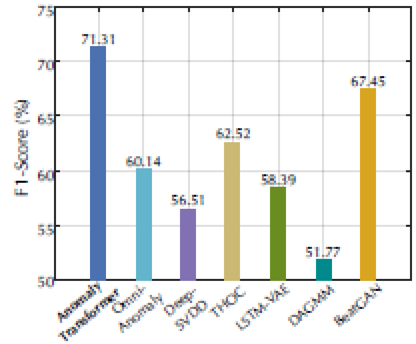
- **Real-world datasets**

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
OCSVM	44.34	76.72	56.19	59.78	86.87	70.82	53.85	59.07	56.34	45.39	49.22	47.23	62.75	80.89	70.67
IsolationForest	42.31	73.29	53.64	53.94	86.54	66.45	52.39	59.07	55.53	49.29	44.95	47.02	76.09	92.45	83.48
LOF	56.34	39.86	46.68	47.72	85.25	61.18	58.93	56.33	57.60	72.15	65.43	68.62	57.89	90.49	70.61
Deep-SVDD	78.54	79.67	79.10	91.92	76.63	83.58	89.93	56.02	69.04	80.42	84.45	82.39	95.41	86.49	90.73
DAGMM	67.30	49.89	57.30	89.60	63.93	74.62	86.45	56.73	68.51	89.92	57.84	70.40	93.49	70.03	80.08
MMPCACD	71.20	79.28	75.02	81.42	61.31	69.95	88.61	75.84	81.73	82.52	68.29	74.73	76.26	78.35	77.29
VAR	78.35	70.26	74.08	74.68	81.42	77.90	81.38	53.88	64.83	81.59	60.29	69.34	90.71	83.82	87.13
LSTM	78.55	85.28	81.78	85.45	82.50	83.95	89.41	78.13	83.39	86.15	83.27	84.69	76.93	89.64	82.80
CL-MPPCA	82.36	76.07	79.09	73.71	88.54	80.44	86.13	63.16	72.88	76.78	81.50	79.07	56.02	99.93	71.80
ITAD	86.22	73.71	79.48	69.44	84.09	76.07	82.42	66.89	73.85	63.13	52.08	57.08	72.80	64.02	68.13
LSTM-VAE	75.76	90.08	82.30	85.49	79.94	82.62	92.20	67.75	78.10	76.00	89.50	82.20	73.62	89.92	80.96
BeatGAN	72.90	84.09	78.10	89.75	85.42	87.53	92.38	55.85	69.61	64.01	87.46	73.92	90.30	93.84	92.04
OmniAnomaly	83.68	86.82	85.22	89.02	86.37	87.67	92.49	81.99	86.92	81.42	84.30	82.83	88.39	74.46	80.83
InterFusion	87.02	85.43	86.22	81.28	92.70	86.62	89.77	88.52	89.14	80.59	85.58	83.01	83.61	83.45	83.52
THOC	79.76	90.95	84.99	88.45	90.97	89.69	92.06	89.34	90.68	83.94	86.36	85.13	88.14	90.99	89.54
Ours	89.40	95.45	**92.33**	92.09	95.15	**93.59**	94.13	99.40	**96.69**	91.55	96.73	**94.07**	96.91	98.90	**97.89**

4. Experiments

4.1 Main results

- **NeurIPS-TS benchmark**

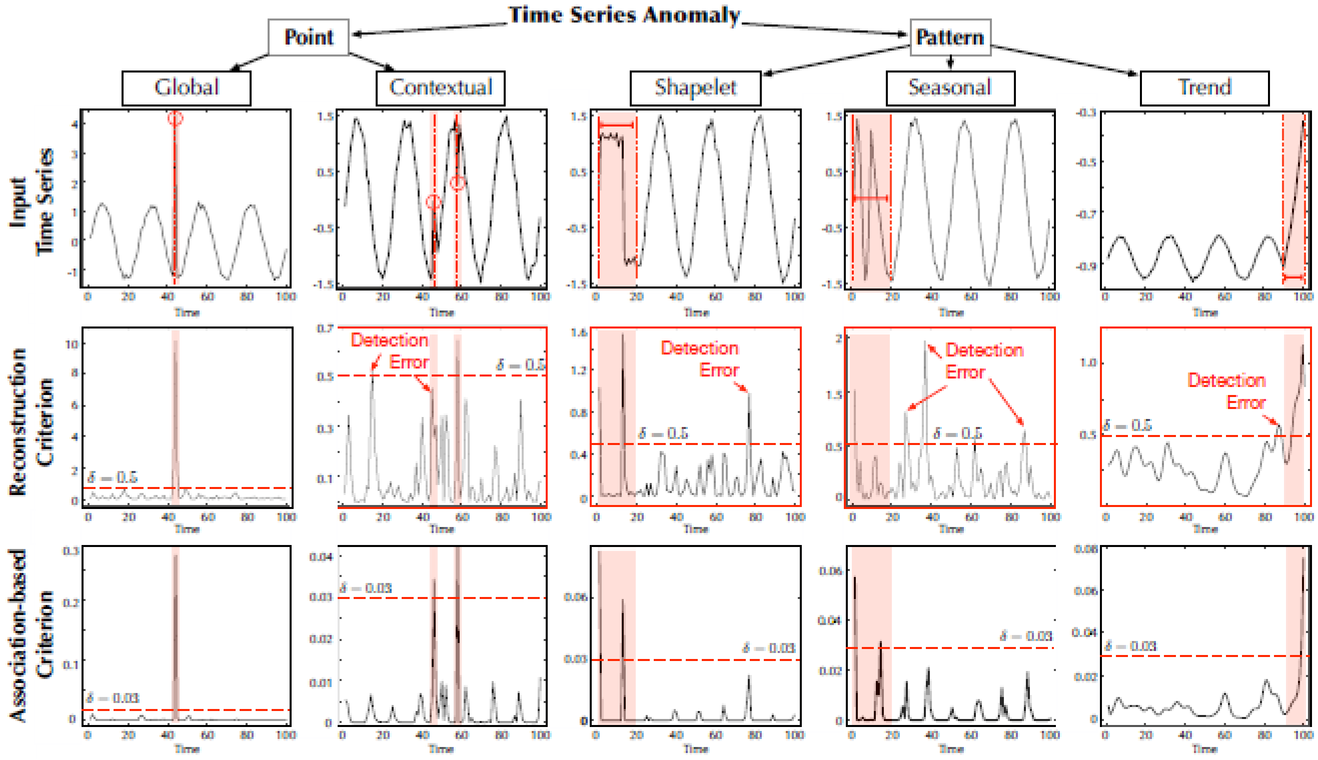


- **Ablation study**

Architecture	Anomaly Criterion	Prior-Association	Optimization Strategy	SMD	MSL	SMAP	SWaT	PSM	Avg F1 (as %)
Transformer	Recon	×	×	79.72	76.64	73.74	74.56	78.43	76.62
Anomaly Transformer	Recon	Learnable	Minmax	71.35	78.61	69.12	81.53	80.40	76.20
	AssDis	Learnable	Minmax	87.57	90.50	90.98	93.21	95.47	91.55
	Assoc	Fix	Max	83.95	82.17	70.65	79.46	79.04	79.05
	Assoc	Learnable	Max	88.88	85.20	87.84	81.65	93.83	87.48
*final	Assoc	Learnable	Minmax	**92.33**	**93.59**	**96.90**	**94.07**	**97.89**	**94.96**

4. Experiments

4.2 Model analysis

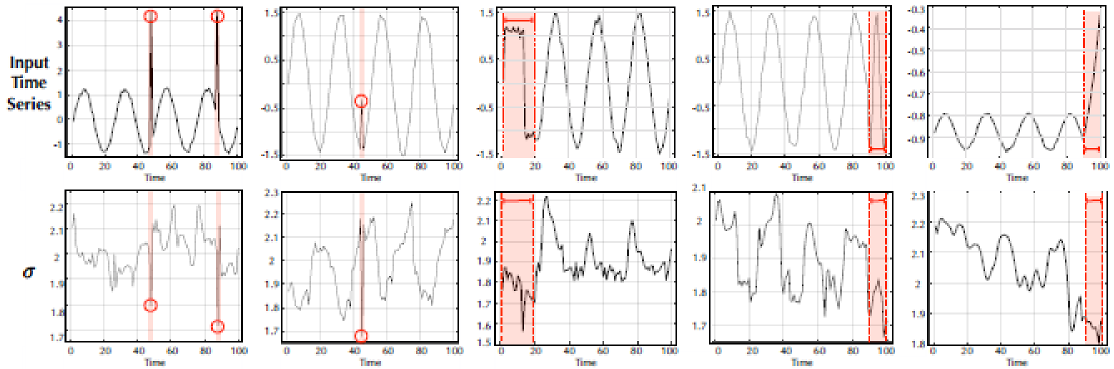


- **Anomaly criterion visualization**
  일반적으로 발생할 수 있는 여러 abnormal 상황에서 detection error가 발생하지 않음.
  (False Positive, False Negative 둘 다)

4. Experiments

4.2 Model analysis

- **Prior-association visualization**



(a) Point-Global (b) Point-Contextual (c) Pattern-Shapelet (d) Pattern-Seasonal (e) Pattern-Trend

- **Optimization strategy analysis**

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Optimization	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours
Abnormal (%)	1.08	0.95	0.86	1.01	0.65	0.35	1.29	1.18	0.70	1.27	0.89	0.37	1.02	0.56	0.29
Normal (%)	0.94	0.75	0.36	1.00	0.59	0.22	1.23	1.09	0.49	1.18	0.78	0.21	0.99	0.54	0.11
Contrast ($\frac{\text{Abnormal}}{\text{Normal}}$)	1.15	1.27	**2.39**	1.01	1.10	**1.59**	1.05	1.08	**1.43**	1.08	1.14	**1.76**	1.03	1.04	**2.64**

adjacent association weights for abnormal and normal time points

감사합니다.