# Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy

| | |
|---|---|
| Author | Jiehui Xu et al. |
| read date | @2022년 6월 22일 |
| Journal | ICLR |
| PDF | [2022 ICLR] Anomaly Transformer Time Series Anomaly Detection with Association Discrepancy.pdf |
| Published Date | 2022 |
| detail | |
| keyword | ATTENTION Anomaly Detection MTS Unsupervised |
| link | https://arxiv.org/pdf/2110.02642.pdf |
| status | Finished! |

## ABSTRACT


- Unsupervised AD는 모델이 distinguishable criterion(구별 가능한 기준)을 도출해야 한다.
- 최근 트랜스포머는 pointwise presentation및 pairwise 연관의 통합된 모델링에서 힘을 보여줬다.
- 우리는 각 시점의 self-attention weight distribution가 whole series와 풍부한 연관성을 구현(embody rich association)할 수 있다는 것을 발견했다.
- 핵심 관측치는 이상치가 희귀하기 때문에, 이상 지점에서 전체 시리즈로 사소한 연관성을 구축하기가 매우 어려워서 이상 징후 연관성은 주로 인접한 시점(adjacent time points)에 집중되어야 한다.
- adjacent-concentration bias(인접 집중 편향)은 연관성 기반 기준(association-based criterion)을 의미, 우리는 **Association Discrepanc**y(연관 불일치)를 강조한다.
- 제안 모델 : 연관 불일치를 계산하는 Anomaly-Attention mechanism 이용 Anomaly Transformer
- A minimax strategy : **Association Discrepancy**의 정상-비정상 구별 가능성을 증폭 위해 개발
- 3개에서(service monitoring, space & earth exploration, and water treatment) AD에서 비지도 학습 6개 SOTA 결과를 보임


## INTRODUCTION

- 이상치는 적고 정상치는 광대하고, 라벨링 작업은 비싸다. → unsupervised setting
- Unsupervised time series AD
  - 모델은 informative representations을 complex temporal dynamics로 부터 학습해야 함
  - LOF, OC-SVM, SVDD 등 고전적 모델은 temporal information을 고려 X
  - Deep NN가 좋은 성능을 달성
- 기존 anomaly criterion : reconstruction 또는 prediction error

  → point by point 기준 score로, temporal context(시간적 맥락)에 대한 포괄적 설명을 제공 못함

  → Association-based Anomaly Criterion을 제시

- Transfomers
  - each time에 대한 temporal association를 self-attention map에서 얻을 수 있음
  - 모든 시점에 대해 association weights distribution를 제시
    - association distribution는 시계열의 period or trend(주기 또는 추세)와 같은 동적 패턴(dynamic patterns)을 나타내는 시간적 맥락(temporal contex)에 대해 더 유익한 설명 제공 가능

→ association distribution을 “**series-association**”으로 부르겠음!

global이라고 생각하면 쉬움 : 전반적 시계열에 대한 상관관계

- prior-Association
  - 이상치가 전체 시계열과 강한 associations 을 만드는 건 어렵다. (정상패턴이 우세해서)
    - 이상치의 연관성은 연속성 → adjacent time points(인접 시점)에 집중

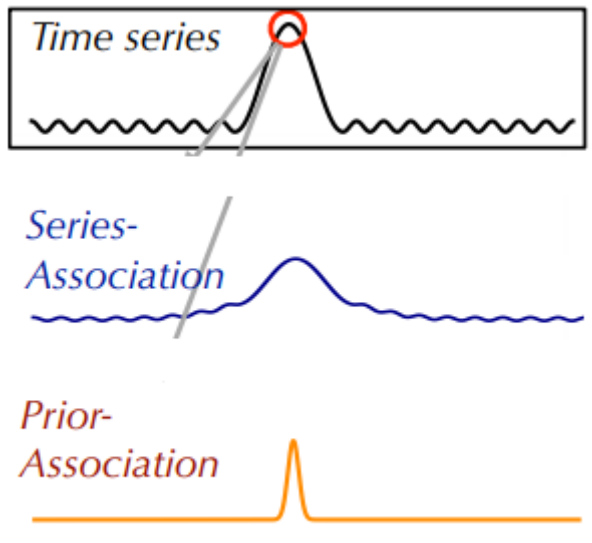
      → adjacent-concentration inductive bias를 “**prior-association**”라고 부르자

      prior을 local이라고 생각하면 쉬움 : 인접한 부분만 보자

- Association Discrepancy
  - 대조적으로 정상 시점은 인접에 제한되지 않고 전체 시계열과 정보적 연관성 발견 가능
    - normal-abnormal distinguishability of the association distribution

      : 연관 분포의 정상-비정상 구별 가능성
    - 즉, prior-association과 series-association 사이 distance = **”Association Discrepancy”**

      → 새로운 anomaly criterion : **”Association Discrepancy(연관 불일치)”**




- **Anomaly Transformer**
  - Anomaly Transformer for association learning
  - Association Discrepancy 계산 위해, Self-attn → **Anomaly-Attention**으로 개선
    - 모델은 각 time point별 prior-association과 series-association 구조로 나눠짐
  - prior-association

    : learnable Gaussian kernel 사용 시점 별 adjacent concentration inductive bias를 제시
  - series-association

    : raw series로부터 학습된 self-attention weights에 해당 됨
  - minimax strategy

    : normal-abnormal distinguishability를 증폭, new association-based criterion 도출 가능

    연관 불일치 값을 효율적으로 추출하기 위한 기법

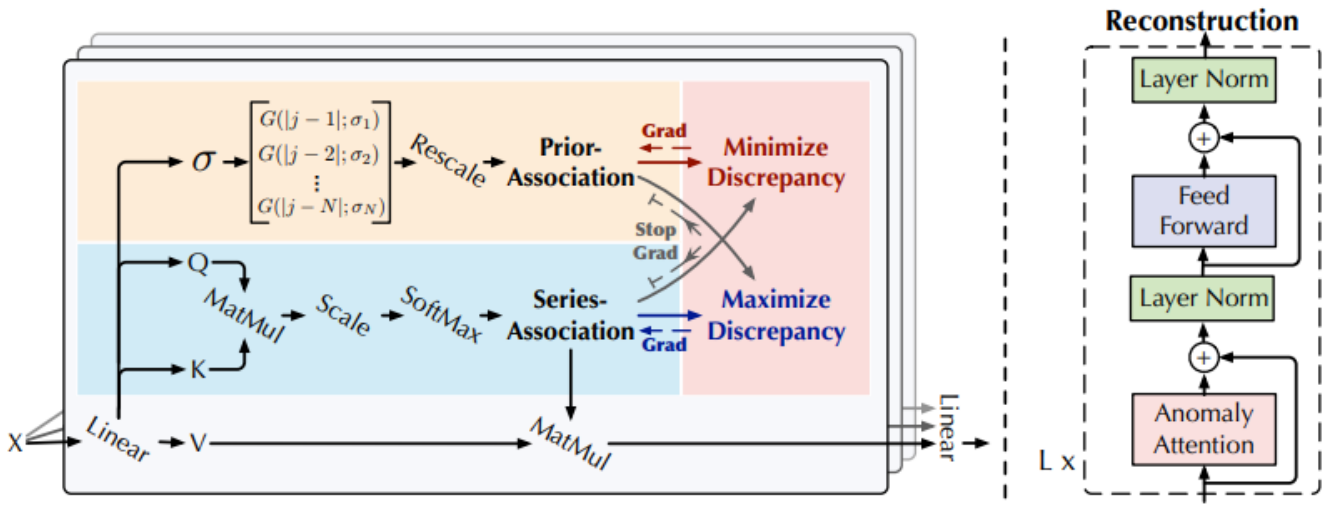
- 기여점
  1. Association Discrepancy를 구현하기 위해 prior-association과 series-association 를 동시에 모델링할 수 있는 Anomaly-Attention mechanism의 **Anomaly Transformer 제안**

2. **Association Discrepancy**의 normal-abnormal 구별가능성을 증폭시키고, new association-based detection criterion를 도출 (최초)
3. Anomaly Transformer는 3개의 실제 적용에 대한 6개의 벤치마크에서 sota 달성

# METHOD

## 3.1 ANOMALY TRANSFORMER

### Overall Architecture : Anomaly Attention | 한 개 layer




- Anomaly-Attention blocks과 feed-forward layers를 번갈아서 쌓음

  : 이러한 stacking 구조는 deep multi-level features에서 underlying association 학습에 도움

- Notation

  input data $\chi = \{x_1, x_2 \cdots, x_N\}, x_t \in R^d, \chi \in R^{N \text{ x } d}$, $l$-th layer

$$\mathcal{Z}^l = \text{Layer-Norm}\left(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\right)$$
$$\mathcal{X}^l = \text{Layer-Norm}\left(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\right),$$

  - $\chi^0$ = Embedding$(\chi)$
    - $\chi^0$ = token(input c를 d_model로 mapping) + positional(위치 정보)임베딩 수행 됨
  - $\chi^{l-1}$= 전 layer의 Output
  - Anomaly-Attention은 Association Discrepancy를 계산 함

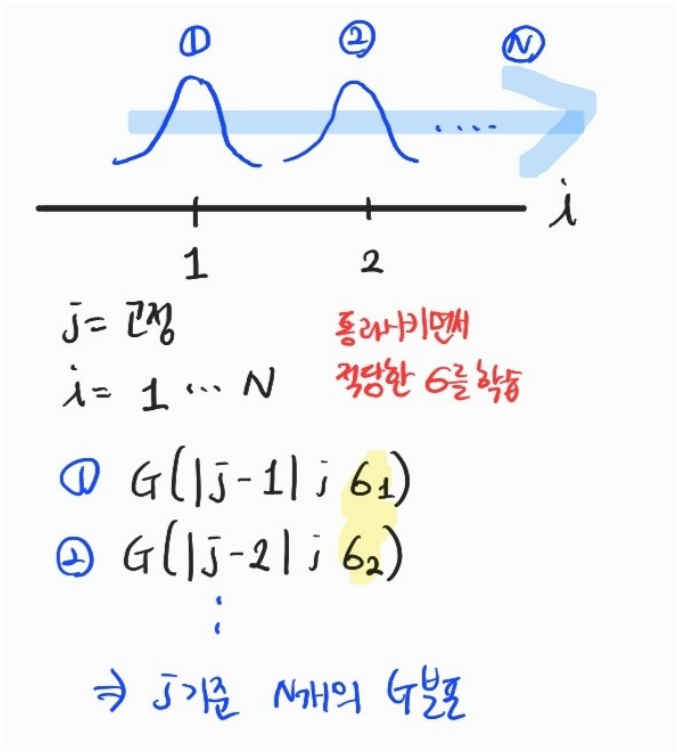
Anomaly-Attention in the l-th layer

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

- prior-association

  : 가우시안 분포를 이용하여 relative temporal distance에 대한 prior(=local)값을 계산

- prior-associations을 다양한 time series patterns에 적용 시킴
- learnable scale parameter σ를 가우시안 커널을 위해 사용




$$\begin{bmatrix} G(|j-1|;\sigma_1) \\ G(|j-2|;\sigma_2) \\ \vdots \\ G(|j-N|;\sigma_N) \end{bmatrix}$$

$$G(|j-i|;\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)$$

**Input:** $\mathcal{X} \in \mathbb{R}^{N\times d_{\text{model}}}$: input; $\mathcal{D} = \left((j-i)^2\right)_{i,j\in\{1,\cdots,N\}} \in \mathbb{R}^{N\times N}$: relative distance matrix
**Layer params:** $\texttt{MLP}_{\texttt{input}}$: linear projector for input; $\texttt{MLP}_{\texttt{output}}$: linear projector for output

$\mathcal{Q},\mathcal{K},\mathcal{V},\sigma = \texttt{Split}\left(\texttt{MLP}_{\texttt{input}}(\mathcal{X}),\texttt{dim=1}\right)$ ▷ $\mathcal{Q},\mathcal{K},\mathcal{V}\in\mathbb{R}^{N\times d_{\text{model}}}, \sigma\in\mathbb{R}^{N\times h}$
**for** $(\mathcal{Q}_m,\mathcal{K}_m,\mathcal{V}_m,\sigma_m)$ **in** $(\mathcal{Q},\mathcal{K},\mathcal{V},\sigma)$**:** ▷ $\mathcal{Q}_m,\mathcal{K}_m,\mathcal{V}_m\in\mathbb{R}^{N\times\frac{d_{\text{model}}}{h}}, \sigma_m\in\mathbb{R}^{N\times 1}$
  $\sigma_m = \texttt{Broadcast}(\sigma_m,\texttt{dim=1})$ ▷ $\sigma_m\in\mathbb{R}^{N\times N}$
  $\mathcal{P}_m = \frac{1}{\sqrt{2\pi}\sigma_m}\texttt{exp}\left(-\frac{\mathcal{D}}{2\sigma_m^2}\right)$ ▷ $\mathcal{P}_m\in\mathbb{R}^{N\times N}$
  $\mathcal{P}_m = \mathcal{P}_m/\texttt{Broadcast}\left(\texttt{Sum}(\mathcal{P}_m,\texttt{dim=1})\right)$ ▷ Rescaled $\mathcal{P}_m\in\mathbb{R}^{N\times N}$
  $\mathcal{S}_m = \texttt{Softmax}\left(\sqrt{\frac{h}{d_{\text{model}}}}\mathcal{Q}_m\mathcal{K}_m^{\text{T}}\right)$ ▷ $\mathcal{S}_m\in\mathbb{R}^{N\times N}$
  $\widehat{\mathcal{Z}}_m = \mathcal{S}_m\mathcal{V}_m$ ▷ $\widehat{\mathcal{Z}}_m\in\mathbb{R}^{N\times\frac{d_{\text{model}}}{h}}$
$\widehat{\mathcal{Z}} = \texttt{MLP}_{\texttt{output}}\left(\texttt{Concat}([\widehat{\mathcal{Z}}_1,\cdots,\widehat{\mathcal{Z}}_h],\texttt{dim=1})\right)$ ▷ $\widehat{\mathcal{Z}}\in\mathbb{R}^{N\times d_{\text{model}}}$
**Return** $\widehat{\mathcal{Z}}$ ▷ Keep the $\mathcal{P}_m$ and $\mathcal{S}_m, m=1,\cdots,h$

- series-association

  : raw series로 부터 전체 시계열에 대한 연관성을 학습 (기존 Transformer 구조와 동일)

  - 얻어진 Attention map을 Softmax를 통해 정규화 → 각 $S^l$의 열은 이산 분포 형식
    - $S_m$ = attention distribution(ex. 0.2, 0.4 · · · ), $\hat{Z}_m$ = attention value

**Input:** $\mathcal{X} \in \mathbb{R}^{N\times d_{\text{model}}}$: input; $\mathcal{D} = \left((j-i)^2\right)_{i,j\in\{1,\cdots,N\}} \in \mathbb{R}^{N\times N}$: relative distance matrix
**Layer params:** $\texttt{MLP}_{\texttt{input}}$: linear projector for input; $\texttt{MLP}_{\texttt{output}}$: linear projector for output

$\mathcal{Q},\mathcal{K},\mathcal{V},\sigma = \texttt{Split}\left(\texttt{MLP}_{\texttt{input}}(\mathcal{X}),\texttt{dim=1}\right)$ ▷ $\mathcal{Q},\mathcal{K},\mathcal{V}\in\mathbb{R}^{N\times d_{\text{model}}}, \sigma\in\mathbb{R}^{N\times h}$
**for** $(\mathcal{Q}_m,\mathcal{K}_m,\mathcal{V}_m,\sigma_m)$ **in** $(\mathcal{Q},\mathcal{K},\mathcal{V},\sigma)$**:** ▷ $\mathcal{Q}_m,\mathcal{K}_m,\mathcal{V}_m\in\mathbb{R}^{N\times\frac{d_{\text{model}}}{h}}, \sigma_m\in\mathbb{R}^{N\times 1}$
  $\sigma_m = \texttt{Broadcast}(\sigma_m,\texttt{dim=1})$ ▷ $\sigma_m\in\mathbb{R}^{N\times N}$
  $\mathcal{P}_m = \frac{1}{\sqrt{2\pi}\sigma_m}\texttt{exp}\left(-\frac{\mathcal{D}}{2\sigma_m^2}\right)$ ▷ $\mathcal{P}_m\in\mathbb{R}^{N\times N}$
  $\mathcal{P}_m = \mathcal{P}_m/\texttt{Broadcast}\left(\texttt{Sum}(\mathcal{P}_m,\texttt{dim=1})\right)$ ▷ Rescaled $\mathcal{P}_m\in\mathbb{R}^{N\times N}$
  $\mathcal{S}_m = \texttt{Softmax}\left(\sqrt{\frac{h}{d_{\text{model}}}}\mathcal{Q}_m\mathcal{K}_m^{\text{T}}\right)$ ▷ $\mathcal{S}_m\in\mathbb{R}^{N\times N}$
  $\widehat{\mathcal{Z}}_m = \mathcal{S}_m\mathcal{V}_m$ ▷ $\widehat{\mathcal{Z}}_m\in\mathbb{R}^{N\times\frac{d_{\text{model}}}{h}}$
$\widehat{\mathcal{Z}} = \texttt{MLP}_{\texttt{output}}\left(\texttt{Concat}([\widehat{\mathcal{Z}}_1,\cdots,\widehat{\mathcal{Z}}_h],\texttt{dim=1})\right)$ ▷ $\widehat{\mathcal{Z}}\in\mathbb{R}^{N\times d_{\text{model}}}$
**Return** $\widehat{\mathcal{Z}}$ ▷ Keep the $\mathcal{P}_m$ and $\mathcal{S}_m, m=1,\cdots,h$

→ 위 두 가지 형태는 각 시점의 시간적 의존성을 유지함

- $\hat{Z}$= $l$-th layer Anomaly Attention 이후 hidden representation

- Reconstruction

  : Reconstruction Loss 및 다음 layer input으로 사용 됨

Association Discrepancy (연관성끼리의 유사도)

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

**Algorithm 2** Association Discrepancy $\text{AssDis}(\mathcal{P}, \mathcal{S}; \mathcal{X})$ Calculation (multi-head version).

**Input:** time series length $N$; layers number $L$; heads number $h$; prior-association $\mathcal{P}_{\text{all}} \in \mathbb{R}^{L \times h \times N \times N}$; series-association $\mathcal{S}_{\text{all}} \in \mathbb{R}^{L \times h \times N \times N}$;

1: $\mathcal{P}' = \texttt{Mean}(\mathcal{P}, \texttt{dim=1})$ ▷ $\mathcal{P}' \in \mathbb{R}^{L \times N \times N}$

2: $\mathcal{S}' = \texttt{Mean}(\mathcal{S}, \texttt{dim=1})$ ▷ $\mathcal{S}' \in \mathbb{R}^{L \times N \times N}$

3: $\mathcal{R}' = \text{KL}\Big((\mathcal{P}', \mathcal{S}'), \texttt{dim=-1}\Big) + \text{KL}\Big((\mathcal{S}', \mathcal{P}'), \texttt{dim=-1}\Big)$ ▷ $\mathcal{R}' \in \mathbb{R}^{L \times N}$

4: $\mathcal{R} = \texttt{Mean}(\mathcal{R}', \texttt{dim=0})$ ▷ $\mathcal{R} \in \mathbb{R}^{N \times 1}$

5: **Return** $\mathcal{R}$ ▷ Represent the association discrepancy of each time point

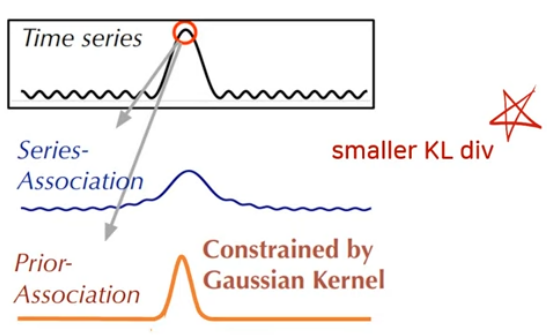
- symmetrized KL divergence

  : 대칭적으로 prior- and series- associations 분포 차이(=유사도)구하는 KL divergence를 사용함

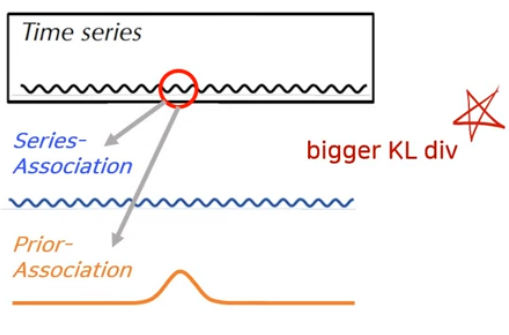
  - 모든 $P^l$과 $S^l$의 row에 대응되는 이산 분포를 KL divergence 계산
  - point-wise별(=each time point) AssDis을 구할 수 있음
  - 이상치는 정상치보다 smaller AssDis를 가지며, 이를 통해 구분할 수 있음

    : 이상치의 경우 $P^l$와 $S^l$가 유사한 분포를 띄지만 → Smaller AssDis

    정상의 경우 가우시안 때문에 $P^l$이 $S^l$ 보다 Peak를 보일 것이다. → bigger AssDis




[비정상 Case]



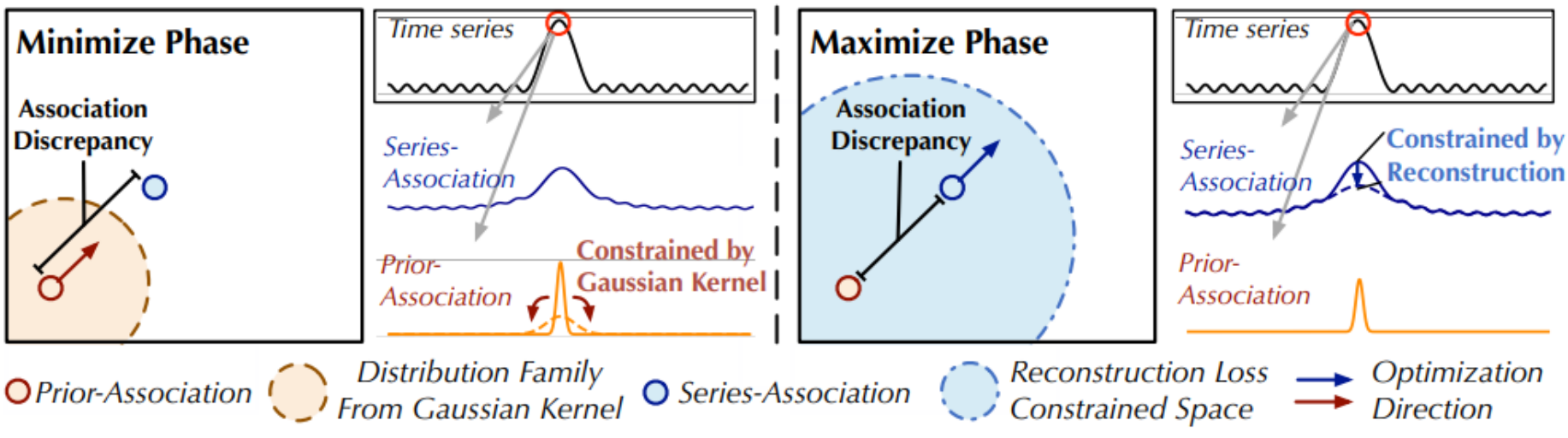

[정상 Case]

## 3.2 MINIMAX ASSOCIATION LEARNING

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

- loss function
  - 왼쪽 식 : $\chi$의 reconstruction loss $\hat{\chi}$
  - 오른쪽 식 : trade off the loss terms으로 AssDis를 Maximize함

    → AssDis를 직접 maximizing 하면 문제 발생 → minimax strategy 전략이 제안 됨

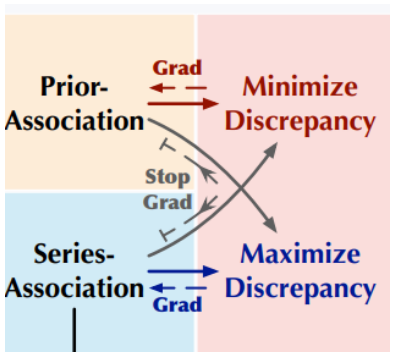



## Minimax Strategy

: AssDis를 직접 maximizing 하면

→ 가우스 커널의 scale 파라미터가 극단적으로 감소하여 prior-association이 무의미해짐

→ minimax strategy 전략이 제안 됨 (detach = stop gradient backpropagation)




$$\text{Minimize Phase: } \mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}, \mathcal{S}_{\text{detach}}, -\lambda; \mathcal{X})$$

$$\text{Maximize Phase: } \mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}}, \mathcal{P}_{\text{detach}}, \mathcal{S}, \lambda; \mathcal{X}),$$

- minimize phase ($S^l$를 detach)

  : $P^l$를 raw series로 부터 학습된 $S^l$에 근사 → $P^l$를 다양한 시간 패턴에 적응하게 함

- maximize phase ($P^l$를 detach)

  : AssDis 최대화를 위해 $S^l$를 최적화, $S^l$이 더 attention을 하게 함

## Association-based Anomaly Criterion

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

: temporal representation(시간 표현) and the distinguishable AssDis를 둘 다 취함

$$\text{AnomalyScore}(\mathcal{X}) \in \mathbb{R}^{N \times 1}$$

: Point-wise AnomalyScore($\chi$)로 lower AssDis → higher anomaly score

# EXPERIMENTS

## dataset

| Benchmarks | Applications | Dimension | Window | #Training | #Validation | #Test (labeled) | AR (Truth) |
|---|---|---|---|---|---|---|---|
| SMD | Server | 38 | 100 | 566,724 | 141,681 | 708,420 | 0.042 |
| PSM | Server | 25 | 100 | 105,984 | 26,497 | 87,841 | 0.278 |
| MSL | Space | 55 | 100 | 46,653 | 11,664 | 73,729 | 0.105 |
| SMAP | Space | 25 | 100 | 108,146 | 27,037 | 427,617 | 0.128 |
| SWaT | Water | 51 | 100 | 396,000 | 99,000 | 449,919 | 0.121 |
| NeurIPS-TS | Various Anomalies | 1 | 100 | 20,000 | 10,000 | 20,000 | 0.018 |

## Implementation

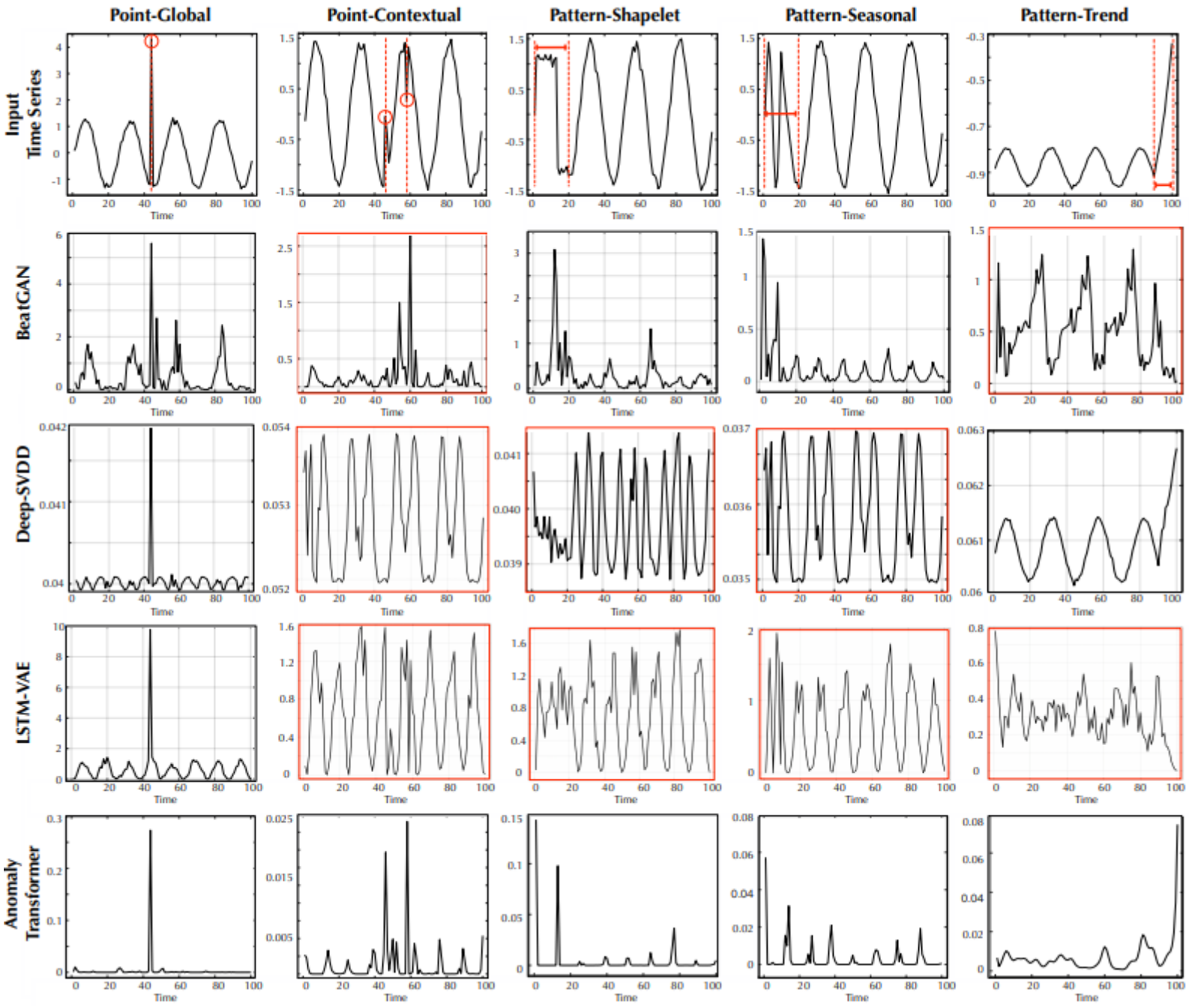
- non-overlapped sliding window to sub-series
- sliding window fixed size = 100
- Anomaly Transformer contains 3 layers
- hidden states $d_{model}$ = 512
- number of heads $h$ = 8
- $\lambda$ = 3 (trade-off two parts of the loss function)

## Result

Table 1: Quantitative results for Anomaly Transformer (*Ours*) in five real-world datasets. The *P*, *R* and *F1* represent the precision, recall and F1-score (as %) respectively. F1-score is the harmonic mean of precision and recall. For these three metrics, a higher value indicates a better performance.

| 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** |

- 성능이 타 모델 대비 뛰어나게 좋은 것을 확인할 수 있다.




- 이상치 유형 별 시각화 (빨간 박스는 탐지에 실패) → Anomaly Transformer는 전부 잡아냄

## 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 |
| | Recon | Learnable | Minmax | 71.35 | 78.61 | 69.12 | 81.53 | 80.40 | 76.20 |
| Anomaly Transformer | 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** |

1. AssDis 사용 : 15.35% 향상
2. Prior-Association 사용 : 8.43% 향상
3. Optimization Strategy 사용 : 7.48% 향상

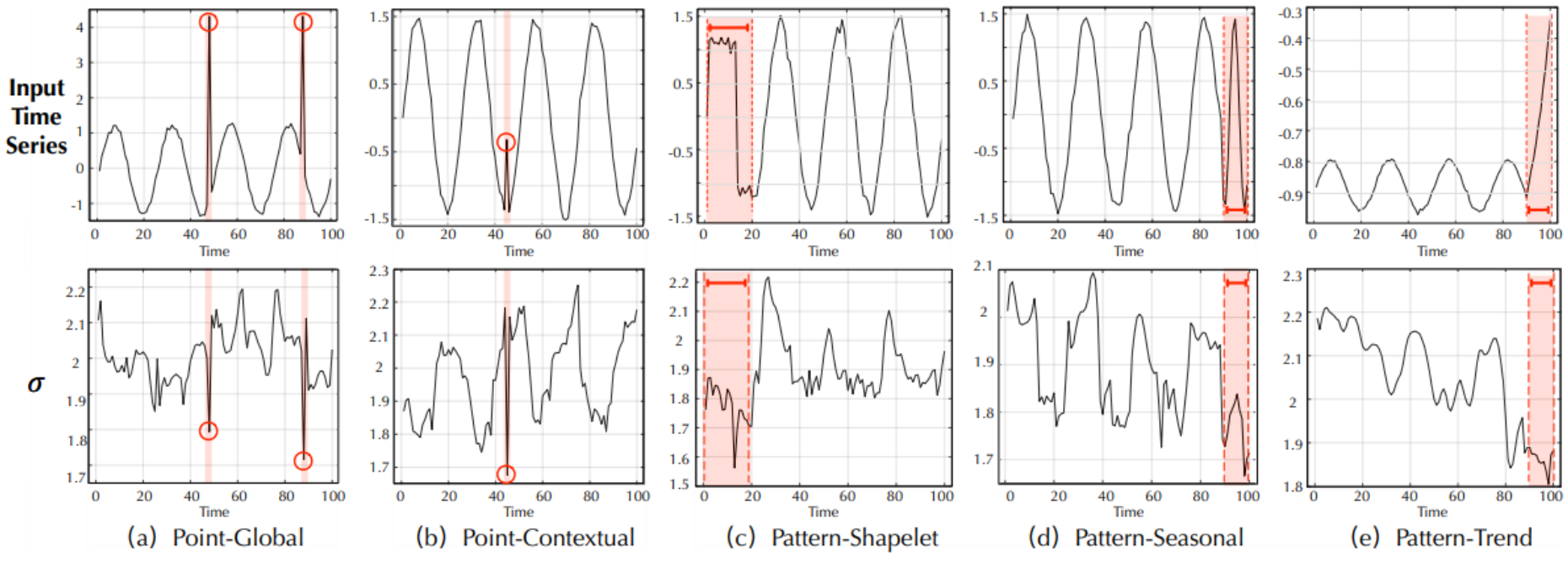



Figure 6: Learned scale parameter $\sigma$ for different types of anomalies (highlight in red).

- 가우시안 분포의 $\sigma$파라미터가 작은 값(=뾰족) 일 때 이상치를 나타낸다. (시그마의 중요도)

## 코드 실행 결과

https://github.com/thuml/Anomaly-Transformer

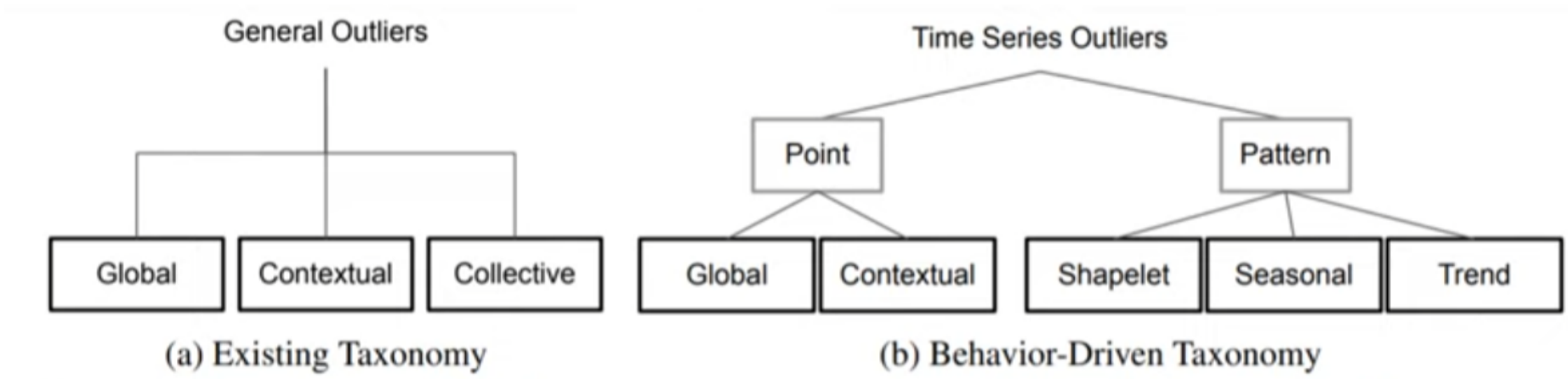
: Dataset - MSL / SMAP / SMD / PSM (SWaT 제외하고 有)

코드도 잘 돌아가고 직관적으로 볼 수 있지만, minmax 전략이 너무 간단하게 보여서 이해가 안감..

```
# Minimax strategy
loss1.backward(retain_graph=True)
self.optimizer.step()
loss2.backward()
self.optimizer.step()
```

---

[참고] 시계열 이상치 유형



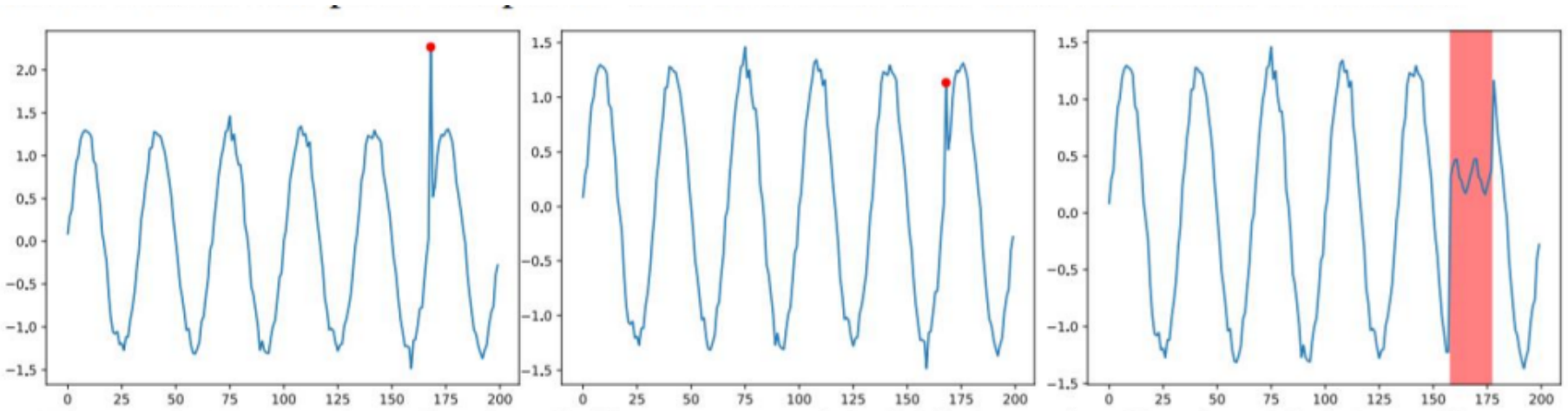
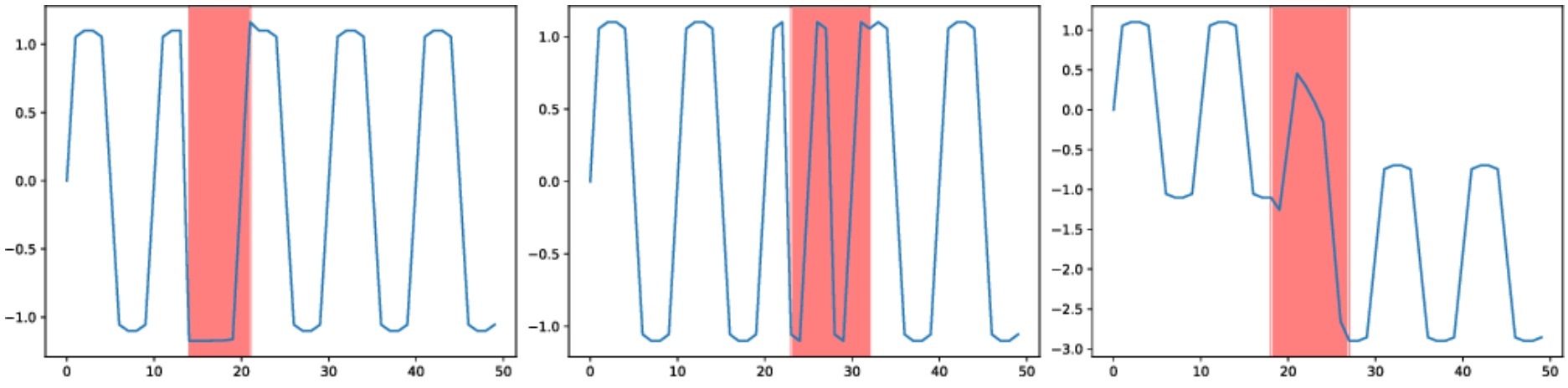



Figure 3: Examples of point (left), contextual (middle), and collective (right) outliers.



(a) Shapelet Outliers (b) Seasonal Outlier (c) Trend Outlier

Figure 4: Illustration of three types of pattern outliers.