ANOMALY TRANSFORMER: TIME SERIES ANOMALY DETECTION WITH ASSOCIATION DISCREPANCY

Jiehui Xu*, Haixu Wu*, Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{xjh20,whx20}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn


Jiehui Xu*


Haixu Wu*


Jianmin Wang


Mingsheng Long

ICLR 2022

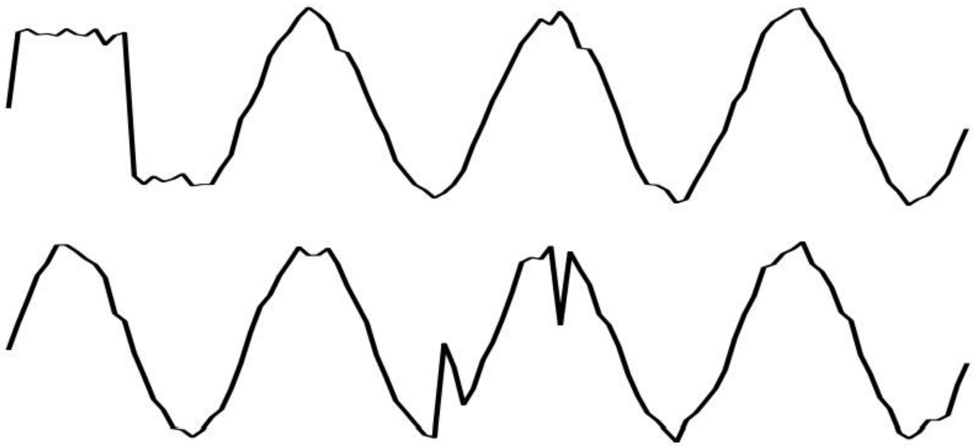
2022.05.23

时间序列异常检测

Service
Monitoring

Space & earth
Exploration

Water
Treatment



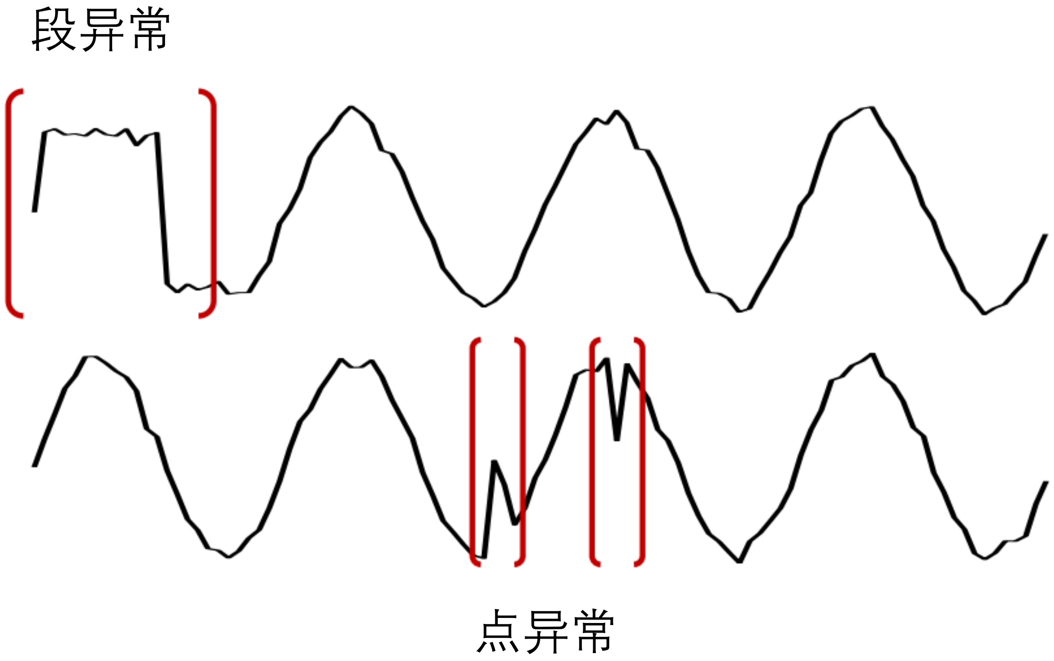
Real-world systems always work continuously and generate successive measurements.

时间序列异常检测

Service
Monitoring

Space & earth
Exploration

Water
Treatment



及时发现故障来保证安全，避免经济损失

检测时间序列中的异常时间点

无监督 时间序列异常检测

分布不平衡 时间点的表示被正常时间点支配

异常通常是罕见的，并隐藏在巨大的正常时间点中，这使得标注困难和昂贵。

无监督任务

阈值判断

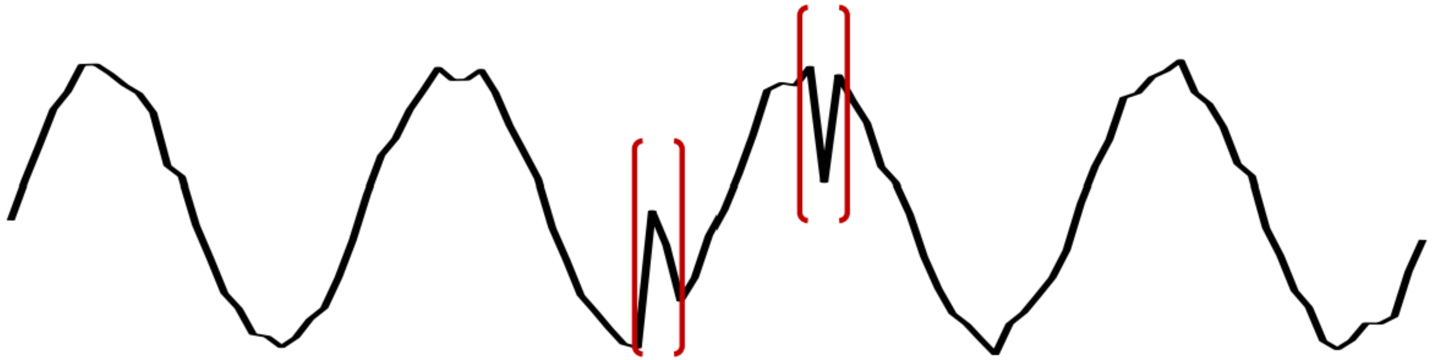
时间序列 → 每个时间点的特征表示 → 异常判据

包含更多的信息量

可区分的

目标：
1. 更好的特征表示
2. 更好的区分判据

相关工作



(1)Classic methods

(e.g. LOF、OC-SVM、SVDD)

- 不考虑时间序列中的时间信息。
- 很难归纳为看不见的真实场景。

每个时间点当作一个数据点，检测离群点。但是忽略了时间模式。
时间序列中最大的特征是时间模式改变。

(2)通过重建和自回归的自监督任务的Recurrent networks

RNN难以捕捉全局long-term 信息

- Point-wise的表示**信息量较少（后面难以区分异常）**，可以由正常点支配。
- 重建或预测错误是逐点的，没有全面的描述。

RNN 考虑了时间信息。重建误差作为判据，异常点重建时候误差大一些。
自回归预测的结果和真实结果差距大时候可能是异常点。

相关工作

(3)显式关联学习

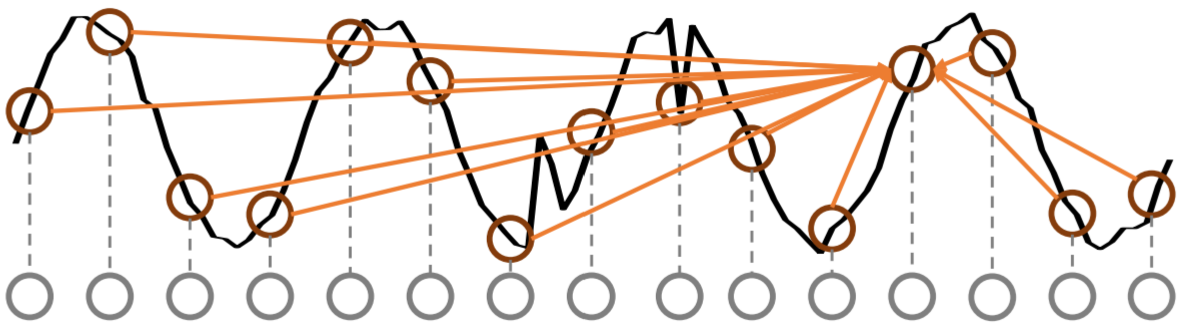
(如：向量自回归、状态空间模型)

- 多变量时间序列的GNN -> 限于单个时间点，对于复杂的时间模式不充分。
- 基于子序列的相似度计算 -> 不能捕获每个时间点的细粒度关联。

比如多变量时间序列，看各个变量间的关系是否变化做异常检测。但是这种关联建模局限在一个时间点上，没有充分利用时间模式。

比如拟合别的段都不同，那么就可能是异常段。但是在细粒度 点的判断要差一些。

Temporal Association（时间关联）



Temporal Association :关联权重分布 到 沿时间维度的所有时间点。

时间上下文的**更多信息量**，表明时间模式，如时间序列的周期或趋势。

比RNN 有更多信息量，比如时间序列具有周期性，关联是以周期形式出现的峰值， 所以 Association 隐含了 时间模式的属性，有更多信息量。

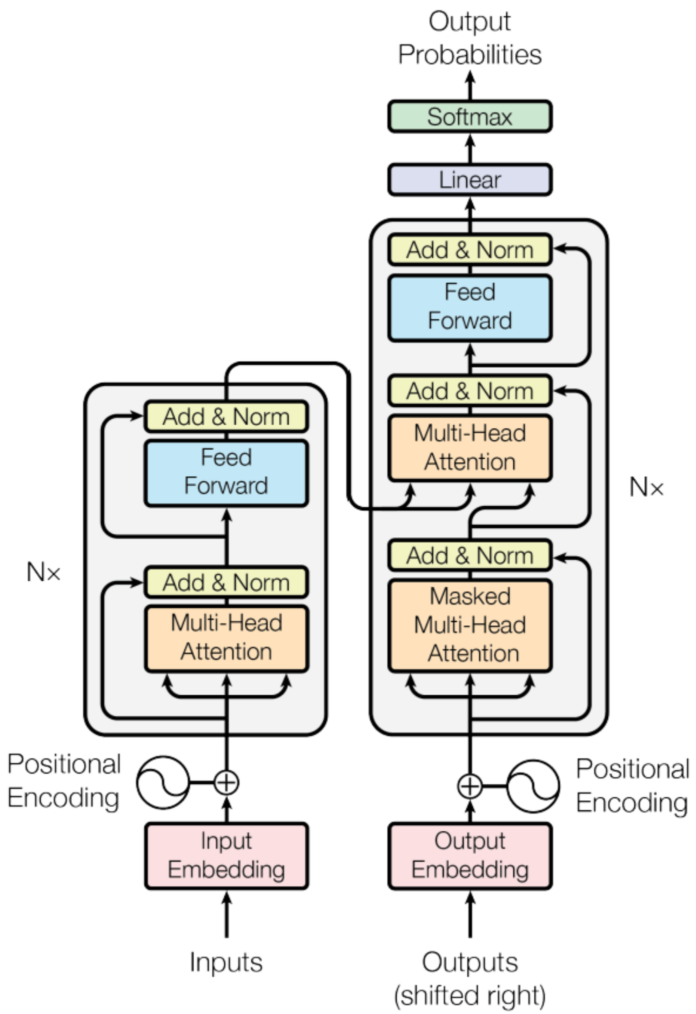
目标：

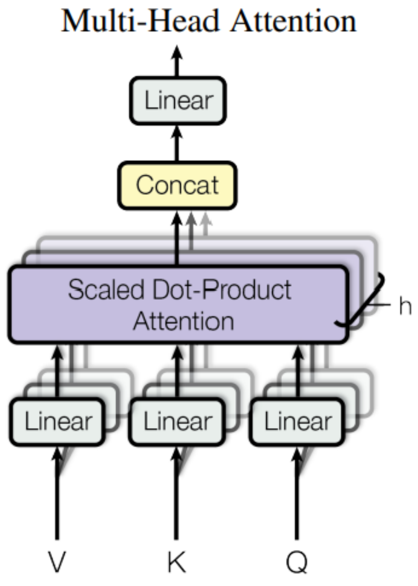
1. 更好的特征表示

2. 更好的区分判据

时间序列不仅可以使用RNN 学到的特征进行表示，**还可以表示为 在当前时间点与整个时间序列的关联权重（ Temporal Association ）。在时间上与其他点怎么连接？ Temporal Association可以表示为一个分布：关联权重的分布。**

Transformer 用于 Series-Association(序列的关联)



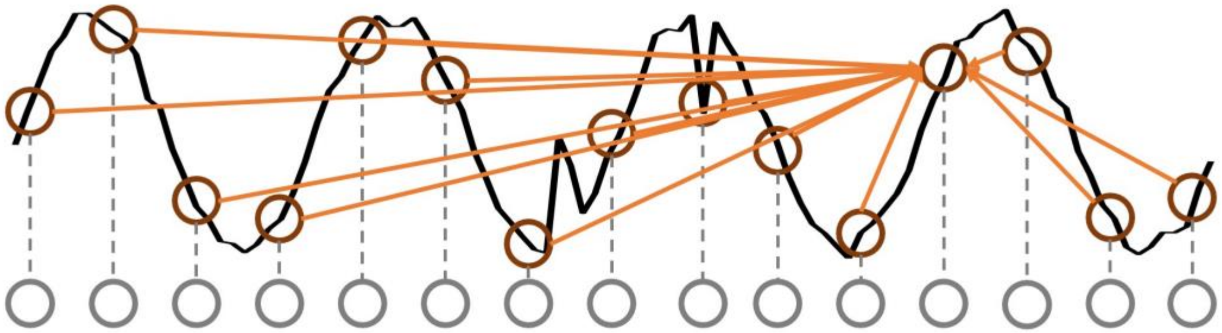


如何挖掘Temporal Association？？

用 Transformer 来建模。
因为是 点级别的 异常检测，所以采用了经典的 self-attention 。

Series-Association
从原始的时间序列中学习的。

Self- attention 图表示当前时刻与其他时刻间关联的大小，就是Temporal Association。定义为：Series-Association。

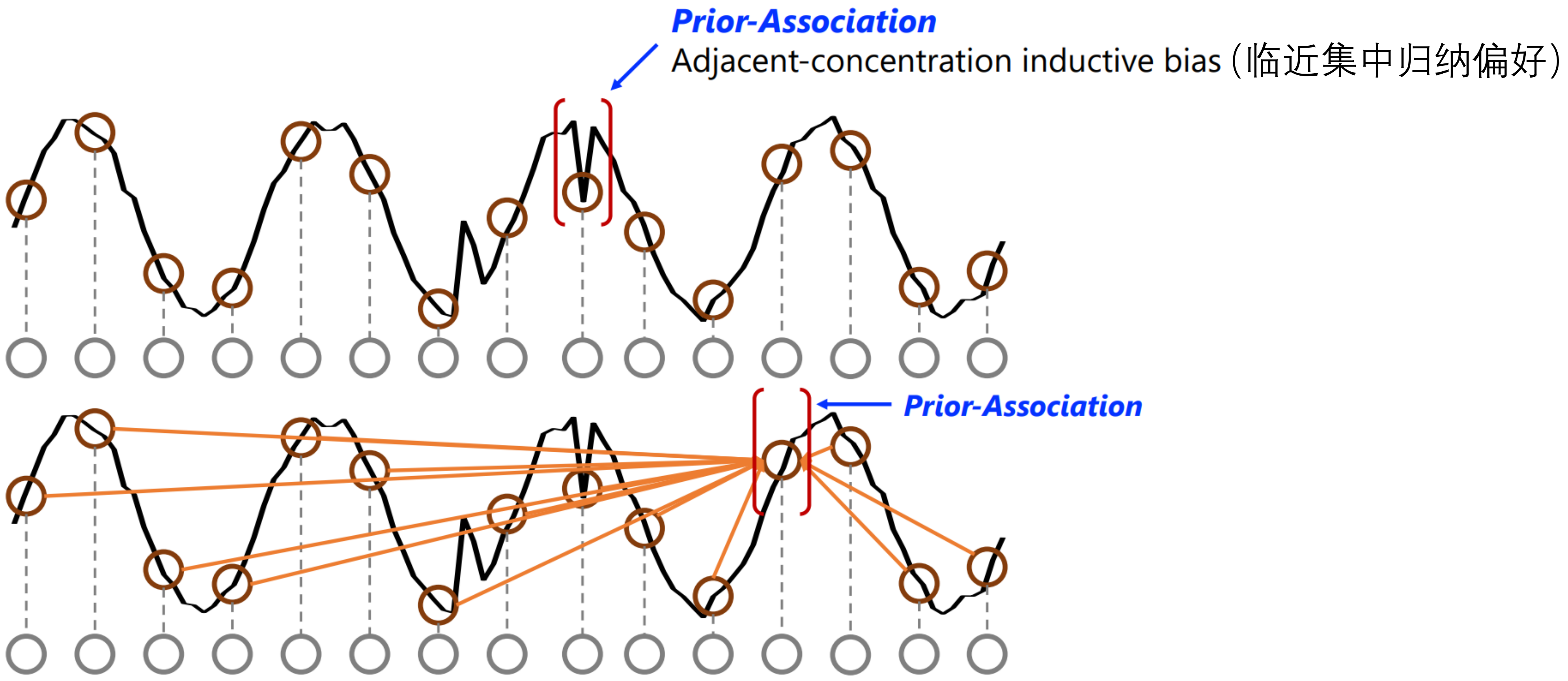


Adjacent-concentration(临近集中) for Prior-Association

时间序列具有连续性，那么当前点和周围的点可能是更相近的。天然有 归纳偏好：临近集中。关联更可能集中在周围。

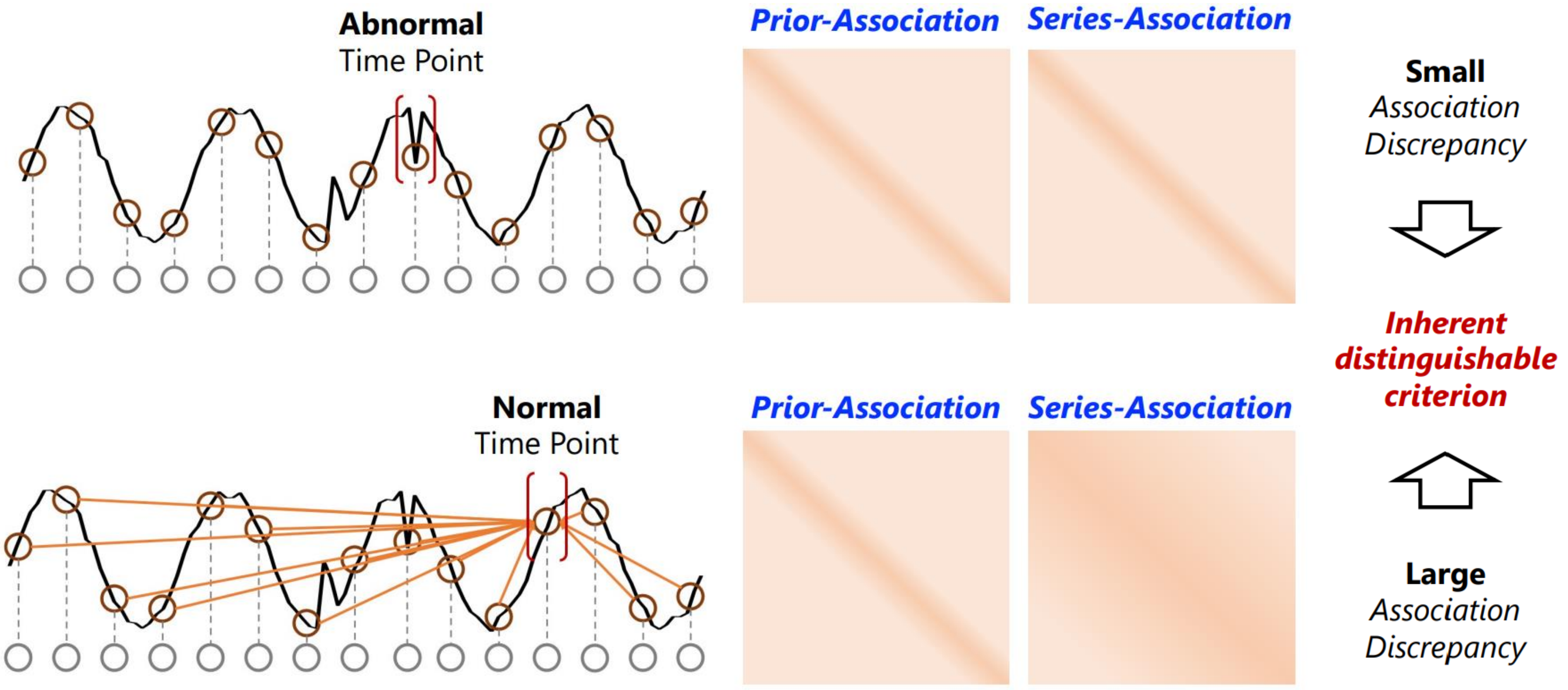
异常点：连续部分是唯一相关部分，临近集中归纳偏好相比与 Series-Association 显著。

正常点： Series-Association比较强，临近集中归纳偏好 和 Series-Association 对比相对弱一些。



Association Discrepancy（关联差异）

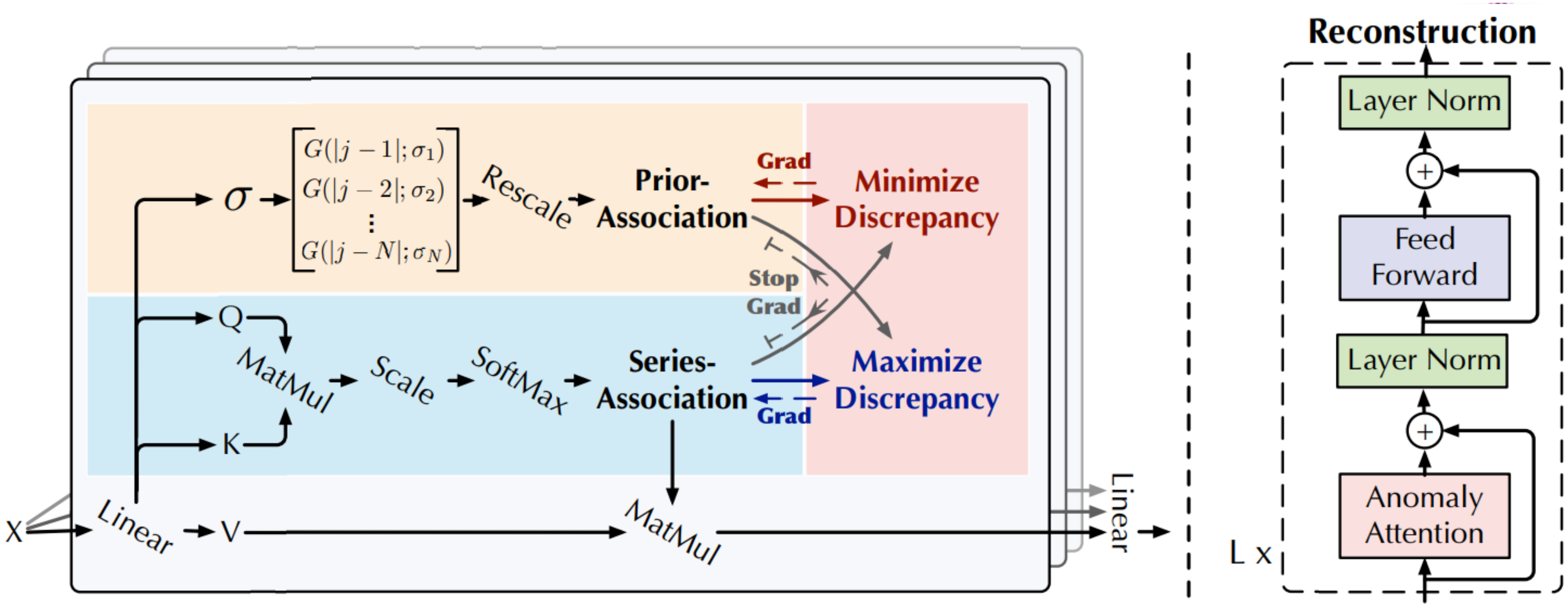
基于以上，得到了一个 判据（Association Discrepancy**）。定义的是：**Prior-Association 和 Series-Association之间的差异。差异原因：异常点难以与整个时间序列都建立起比较强的 Series-Association，那么 Series-Association 就是对角线集中的。 Prior-Association 也是对角线集中的。两个差异小。
正常点的该差异较大。



Association 挖掘总结

发现 **Temporal Association** 信息量更大。又挖掘了Adjacent-concentration inductive bias。合并起来得到一个Association Discrepancy（关联差异）的判据。
完成了上面两个目标：1. 有信息含量的表示，2. 足够判别性的判据。

Anomaly Transformer

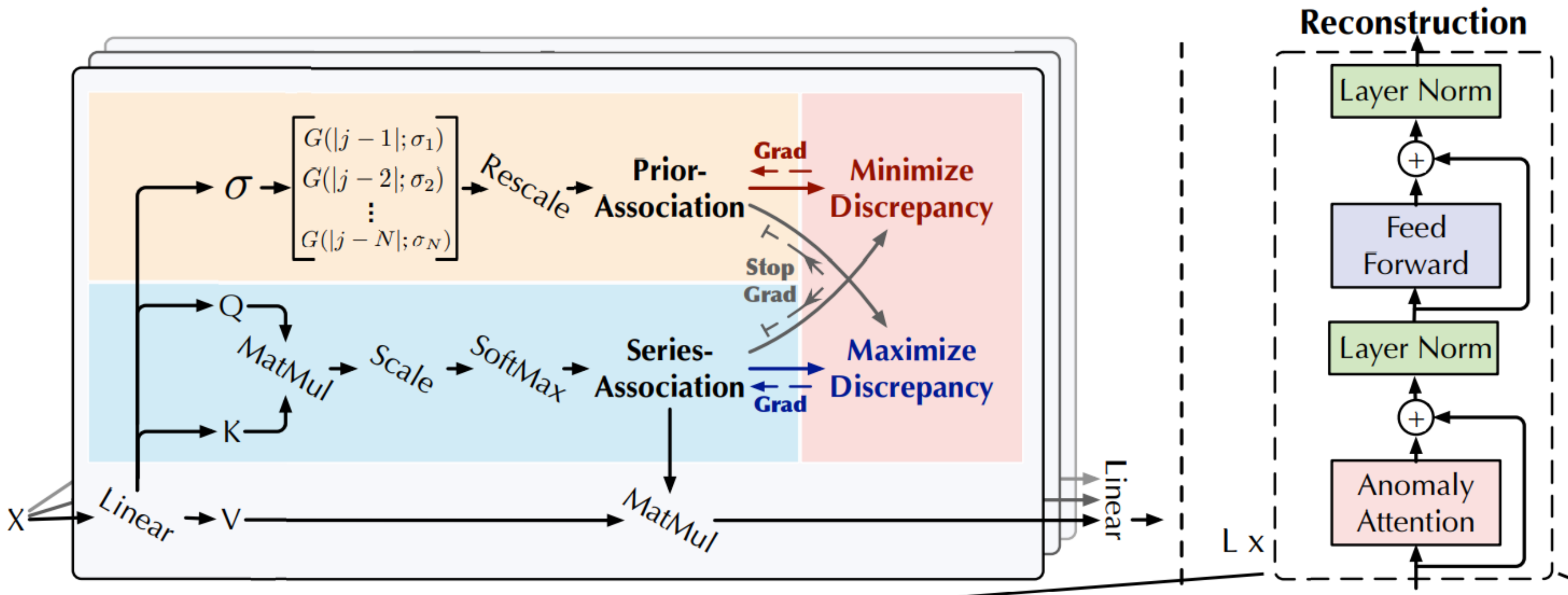


(1)架构:具有 Anomaly-Attention机制 的Anomaly Transformer

(2)训练策略: Minimax 关联学习

(3)准则: 基于关联的异常准则

整体架构



self-attention
替换为 Anomaly Attention

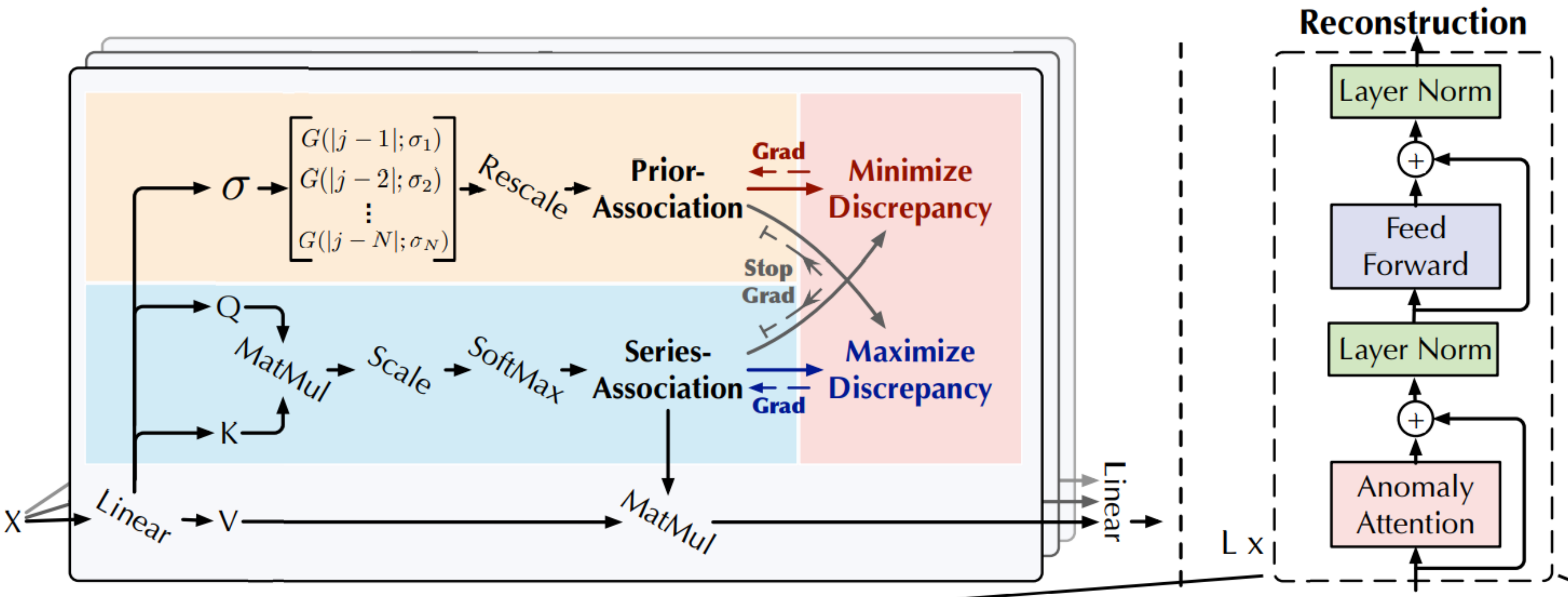
$$\mathcal{Z}^l = \text{Layer-Norm}\Big(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\Big)$$

$$\mathcal{X}^l = \text{Layer-Norm}\Big(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\Big),$$

Learning underlying associations from deep **multi-level** features.

叠加多个模块可学习到多层特征

整体架构



self-attention
替换为 Anomaly Attention

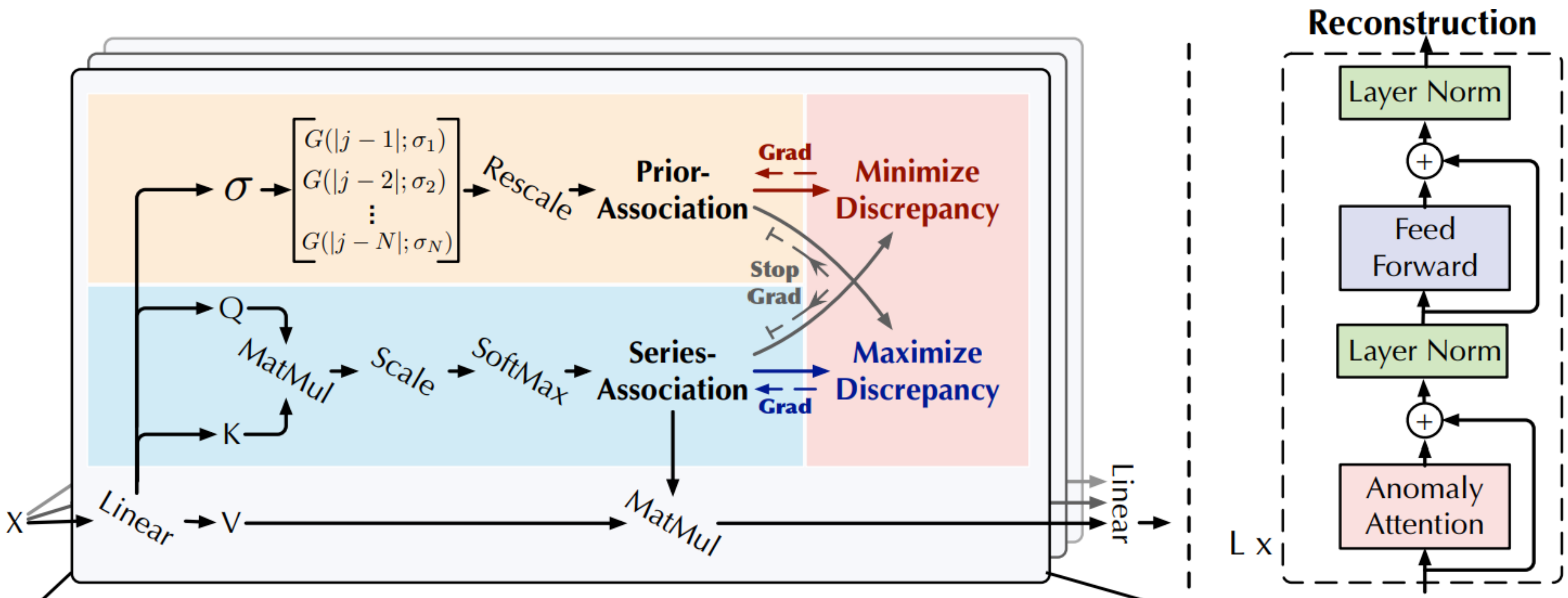
$$\mathcal{Z}^l = \text{Layer-Norm}\Big(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\Big)$$

$$\mathcal{X}^l = \text{Layer-Norm}\Big(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\Big),$$

Learning underlying associations from deep **multi-level** features.

叠加多个模块可学习到多层特征

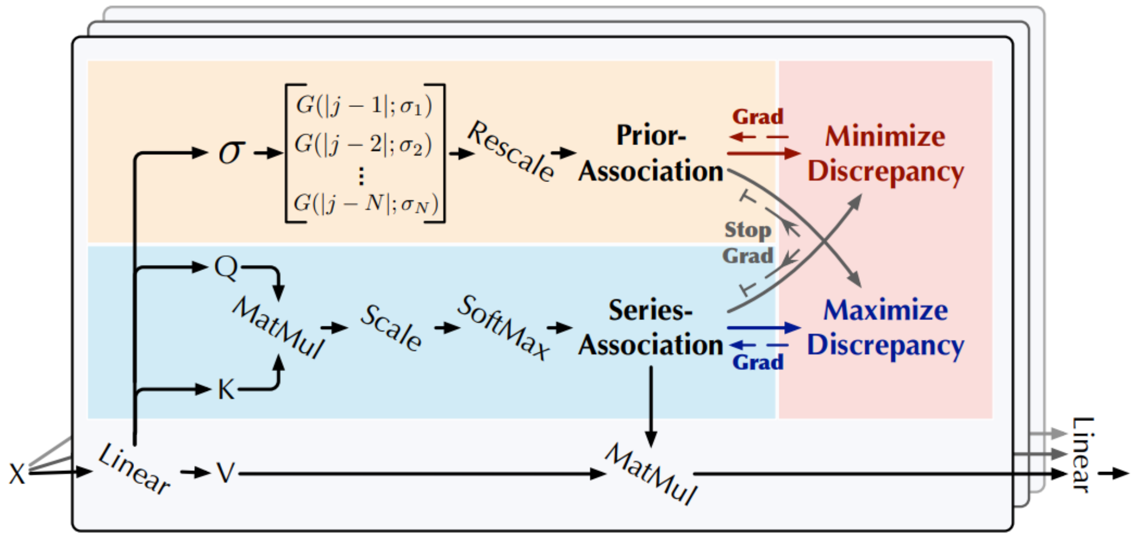
整体架构



Double branches structure to

model the **prior-association** and **series-association** simultaneously.

Anomaly-Attention

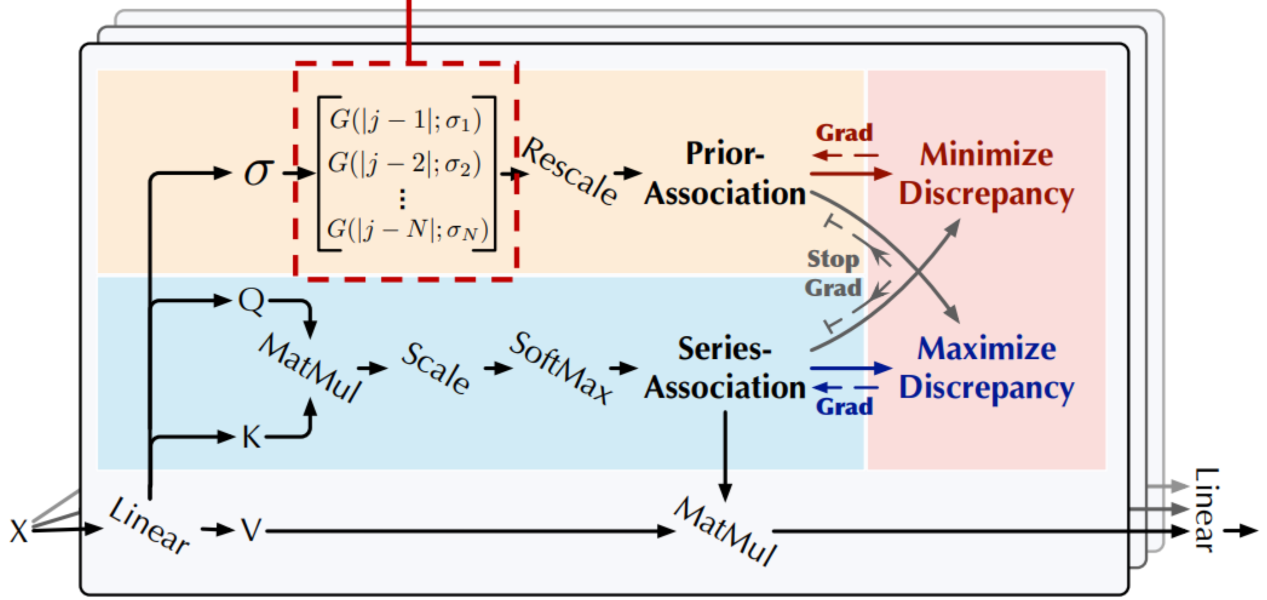


$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

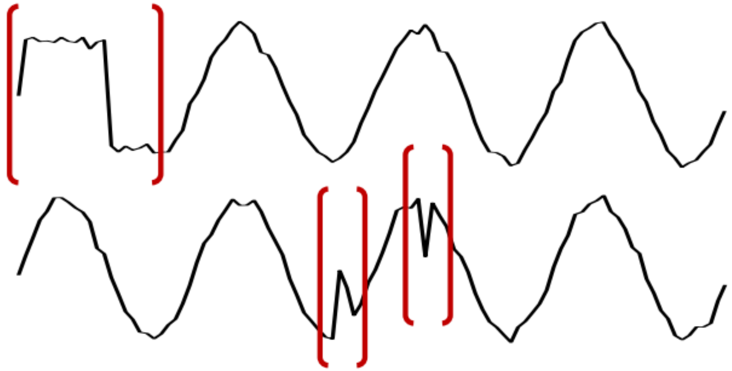
Anomaly-Attention

临近集中的先验，选取了一个可学习的高斯核，学习的是尺度参数 σ。



learnable Gaussian kernel

making prior-associations adapt to the **various time series patterns**

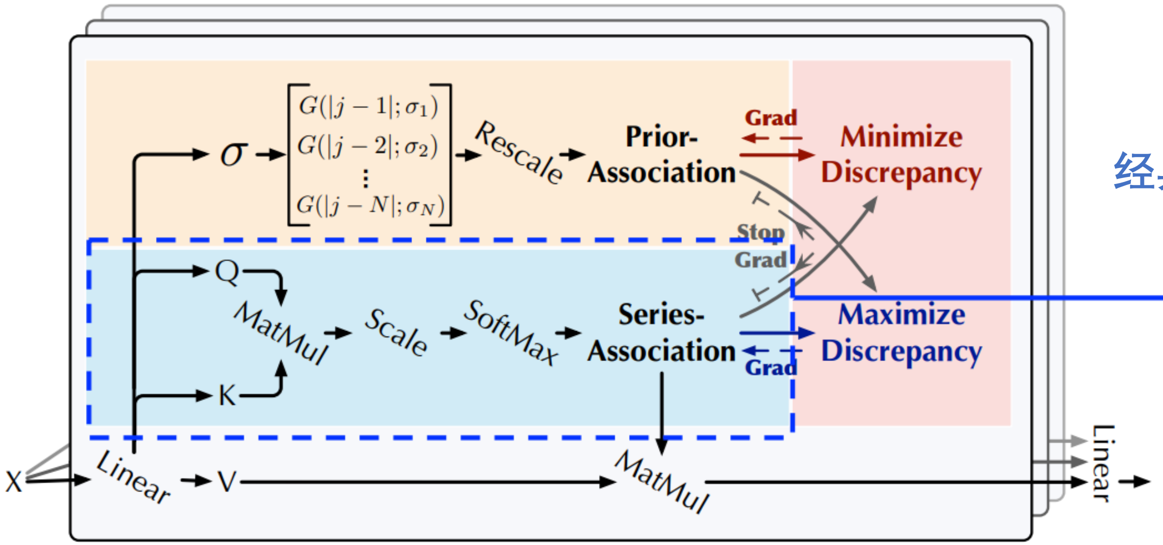


$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{QK}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

尺度参数 σ 使学到的 Prior-Association 适应各种时间模式，
比如段异常， σ 可能大一些。
点异常， σ 可能小一些。
得到一个自适应的 Prior-Association 。

Anomaly-Attention



经典 self- Attention 来做 Series-Association

Self-Attention

Find the most effective associations from raw series

$$\text{Prior-Association: } \mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

$$\text{Series-Association: } \mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

无监督任务，特征表示来自于重建能力，即由重建误差约束的特征学习。为了重建的更好，会尽可能在时间序列内挖掘最有效的时间依赖。self-Attention来找到时间序列内关键的**Temporal Association**。

Association Discrepancy

$$\text{Prior-Association: } \mathcal{P}^l = \mathrm{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right)$$

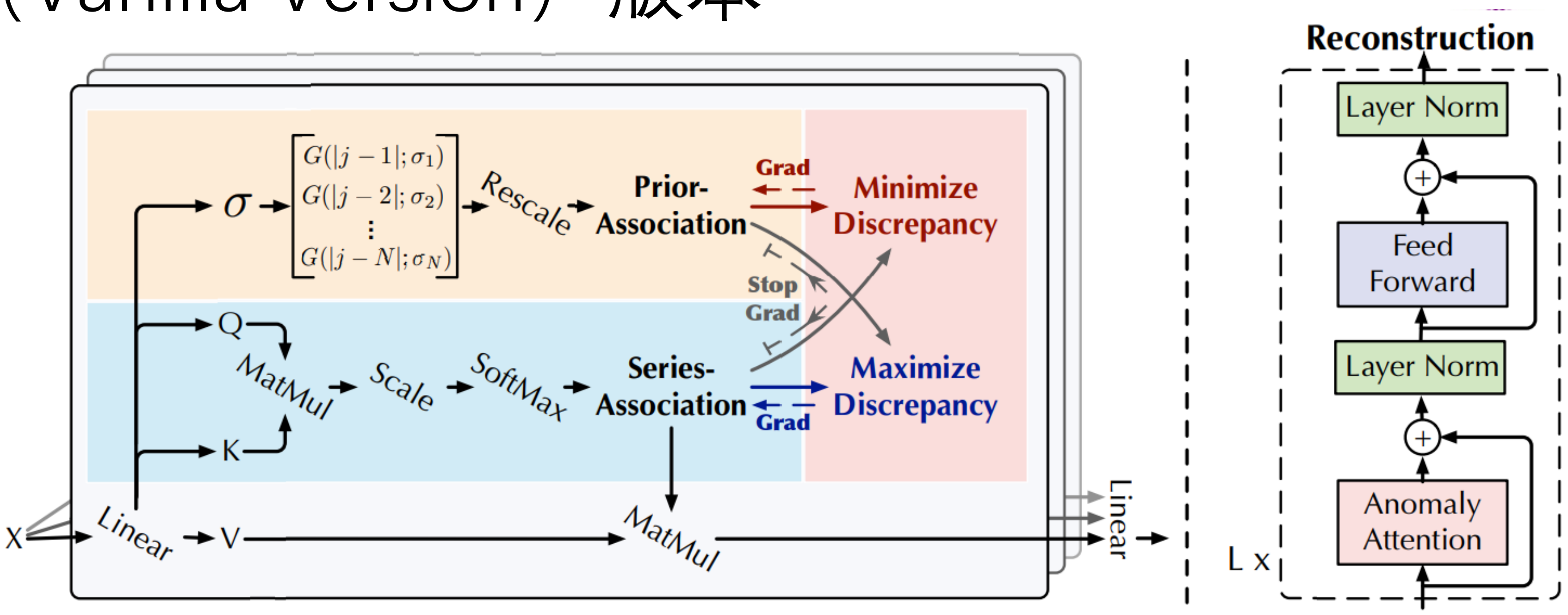
$$\text{Series-Association: } \mathcal{S}^l = \mathrm{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{model}}}}\right)$$

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

对称的 KL 散度 between 多层次 prior- and series-associations (The adjacent-concentration property of series-association)

对称的KL散度计算距离，表示两个分布间的信息增益。

训练策略（Vanilla Version）版本



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

Representation Learning

第一项：重建误差，用于约束模型，学习到一个有意义的表示，如约束序列关联来挖掘时间序列中的有效依赖；
第二项：减去 Association Discrepancy

仅仅使用重建loss 来约束，可能学习到的特征区分性不够，
loss 增加一些项来使得正常点-异常点易分

训练策略（Vanilla Version）

Loss 减小，那么 该项要增大

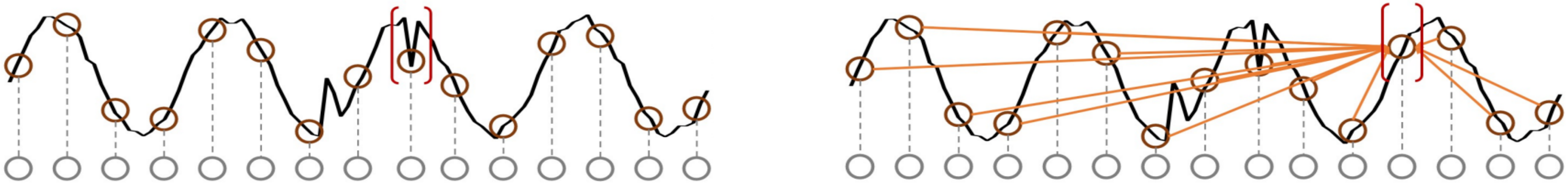
$$\mathcal{L}_{\text{Total}}(\widehat{\mathcal{X}},\mathcal{P},\mathcal{S},\lambda;\mathcal{X}) = \|\mathcal{X}-\widehat{\mathcal{X}}\|_{\mathrm{F}}^{2} - \lambda \times \|\mathrm{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\|_{1}$$

Association Discrepancy增大意味着：迫使每个点更少的关注临近区域；对异常点是致命的，因为异常点无法由其他点重建，它依赖的只有临近点；对正常点造成的损害更少一些，因为它符合 正常模式，整个时间序列的其他部分也符合正常模式，可能它的重建loss 会小一些。

当 Association Discrepancy 增大时，逼迫正常点和异常点之间的差异越来越大。这样训练策略的好处：放大正常点和异常点之间的差别。

异常检测领域的黄金法则：设计出一些困难的任务，正常点和异常点都会变差，正常点变差的幅度小一些，异常点根本不可能完成这个任务，正常点和异常点就可以区分开。

Enlarge the association discrepancy ⟶ Paying less attention to adjacent area



Abnormal time points have the adjacent-concentrate inductive bias ⟶ Making the reconstruction of abnormal time points harder

Amplify the difference between normal and abnormal points

Association discrepancy: the adjacent-concentration property of series-association

训练策略（Vanilla Version）

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

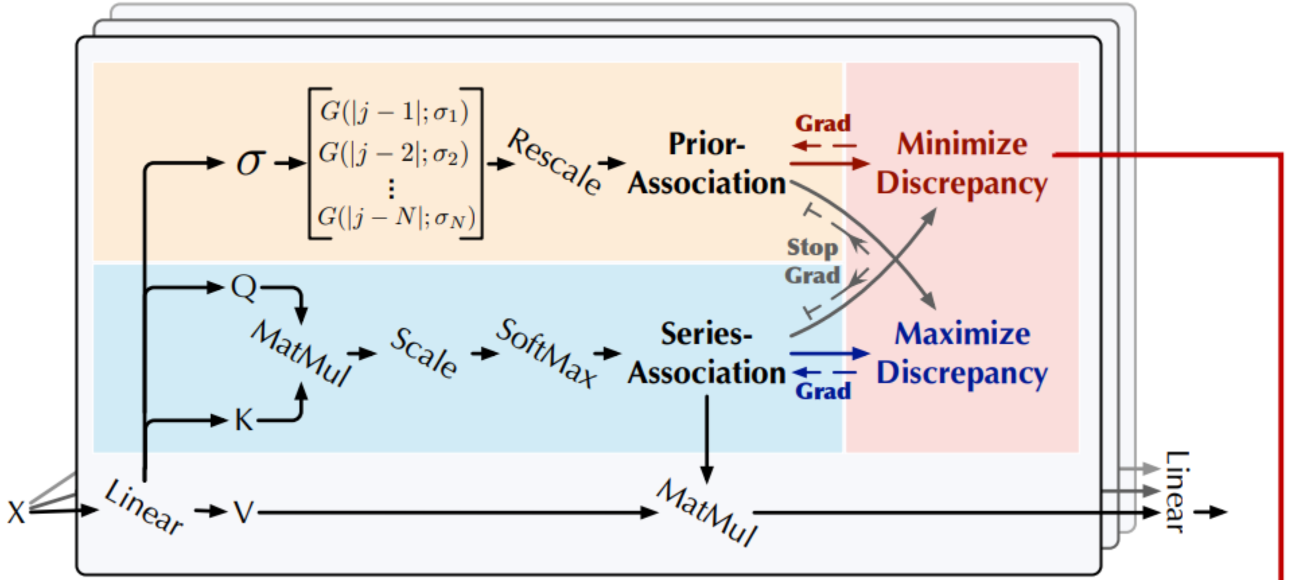
Association Discrepancy 有两个自由度（P，S），让该项变大有个退化版本，只要让高斯核函数的尺度参数 σ 逼近于0 即可，获得了极大值，只学习到了一个退化的特征。

直接最大化 Association Discrepancy将会极大地减少高斯核函数的尺度参数。

Association discrepancy: the adjacent-concentration property of series-association

极大极小关联学习

核心：一次只优化一个特征。



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

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

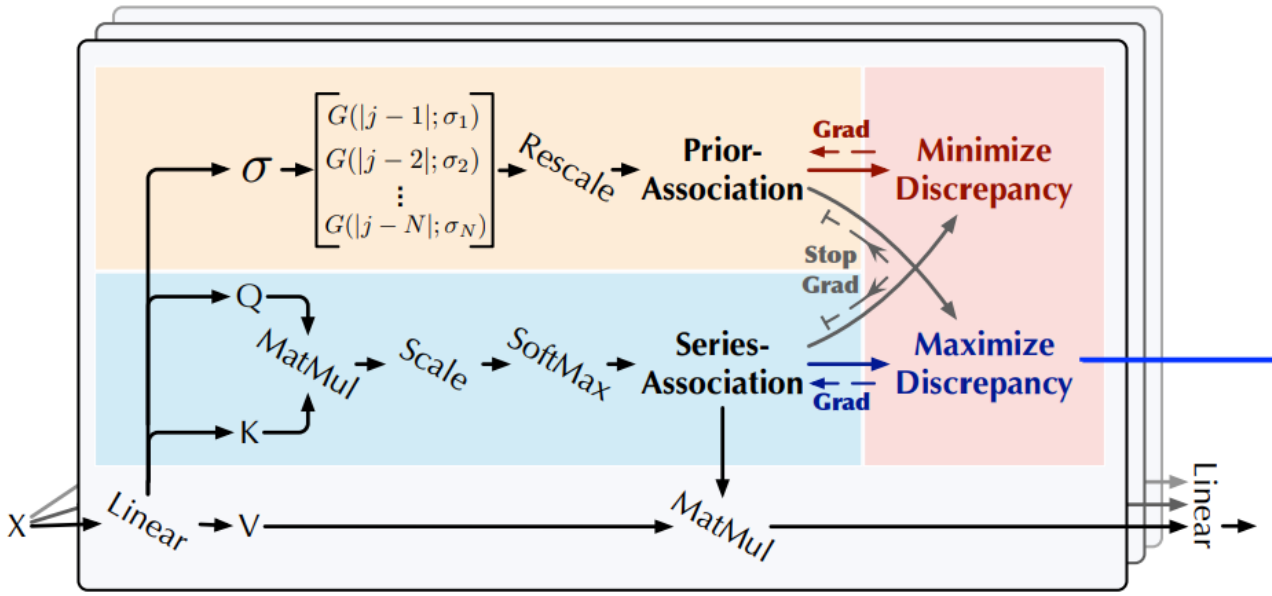
Learning prior-association $\mathcal{P}$ to avoid degeneration

退化

只优化 prior-association， Prior-association 逼近 Series-association，
让学习到的高斯核函数适应不同的时间模式过程。

极大极小关联学习

核心：一次只优化一个特征。



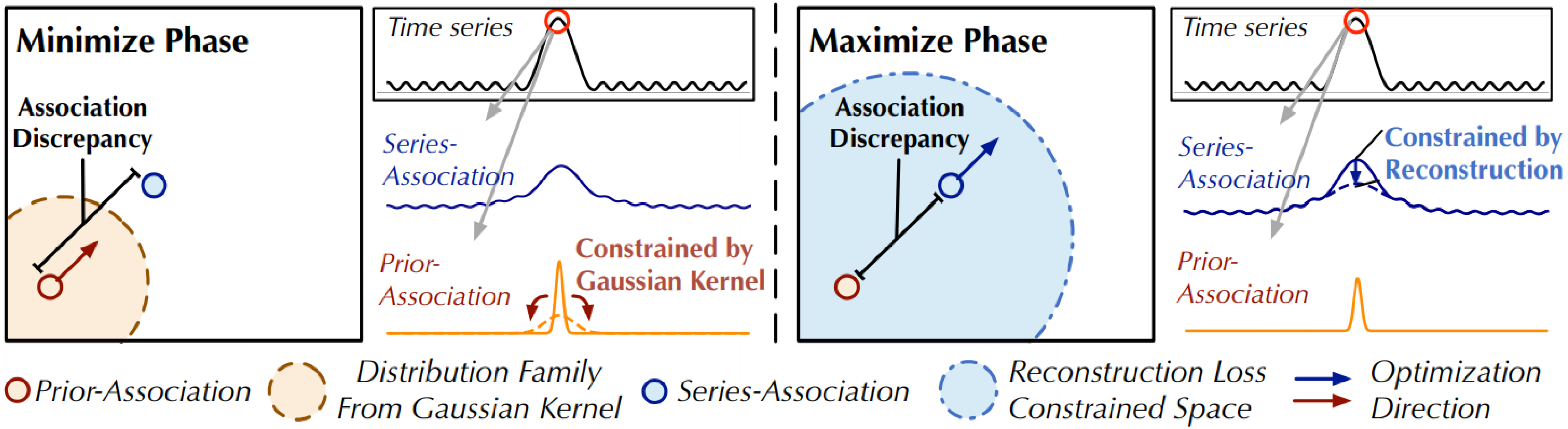
Minimize Phase: $\mathcal{L}_{\mathrm{Total}}(\mathcal{X}, \mathcal{P}, \mathcal{S}_{\mathrm{detach}}, -\lambda; \mathcal{X})$

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

Amplify the difference between normal and abnormal points

只优化Series-association，让Series-association远离 Prior-association ，让学习到的attention 图 对角线的值尽可能小，降低了对临近区域的依赖，可以逼迫异常点更难被重建。

极大极小关联学习 核心：一次只优化一个特征。



1. 获得一个更好的 关联差异的估计
2. 放大异常-正常 可分性。

Association-based Anomaly Criterion

基于关联差异设计了一个 异常判据，卡阈值做异常检测。 点的anomaly scores > 阈值将被检测为异常

$$\mathrm{AnomalyScore}(\mathcal{X}) = \underbrace{\mathrm{Softmax}\Big(-\mathrm{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\Big)}_{\text{归一化关联差异}} \odot \underbrace{\Big[\|\mathcal{X}_{i,:} - \widehat{\mathcal{X}}_{i,:}\|_2^2\Big]}_{\text{重建误差}}{}_{i=1,\cdots,N}$$

相乘的好处：考虑一个异常点，假设重建的好，那重建误差小，那么更关注临近点，那么关联差异就比较大；重建比较差，重建误差就比较大；异常点的两种情况都会使得AnomalyScore 变大，至少一项是比较大的。

Item for Anomalies	(1) Good Reconstruction	(2) Bad Reconstruction
Normalized Association Discrepancy	**Larger**	Unknown
Reconstruction Error	Smaller	**Larger**

相互协作，提高检测性能。

实验

Table 1: Details of benchmarks. AR represents the truth abnormal proportion of the whole dataset.

Benchmarks	Applications	Dimension	Window	#Training	#Validation	#Test	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

三个实际应用的六个基准

主要结果（SOTA 超过 18 个基线）

Classic methods

Density-based

Autoregression-based

Reconstruction-based

Clustering-based

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**

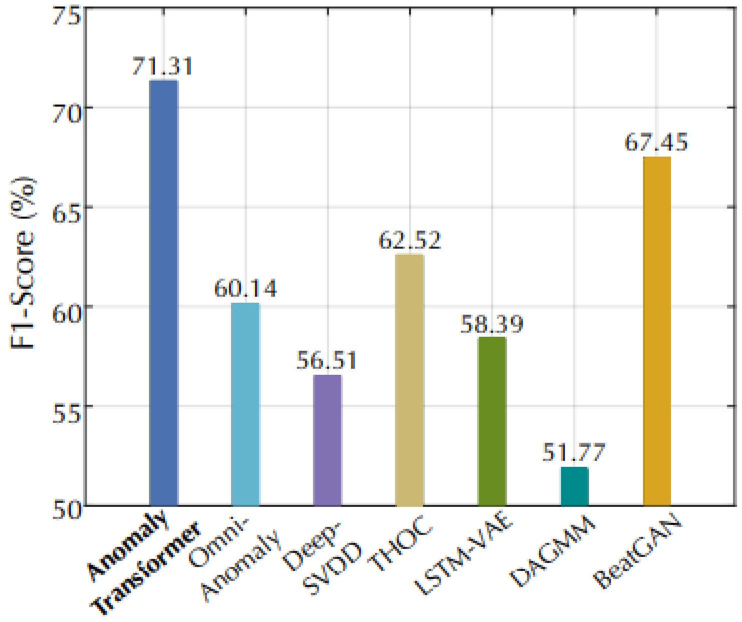
主要结果（SOTA 超过 18 个基线）

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
Previous SOTA	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**

NeurIPS-TS benchmark

- Time series Anomaly
  - Point
    - Global
    - Contextual
  - Pattern
    - Shapelet
    - Seasonal
    - Trend



Achieve SOTA on various anomalies.

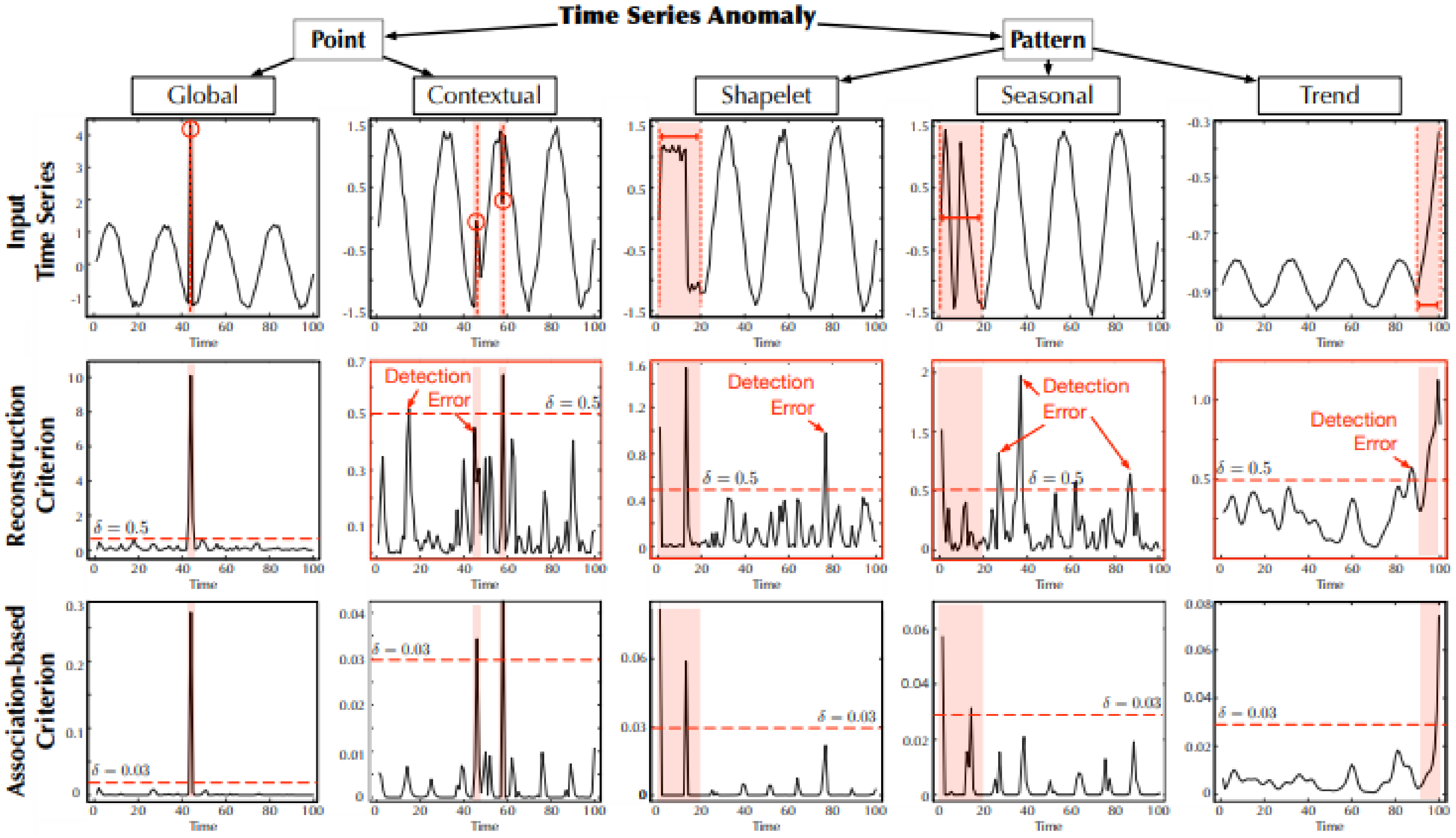
消融研究

Table 2: Ablation results (F1-score) in anomaly criterion, prior-association and optimization strategy. *Recon*, *AssDis* and *Assoc* mean the pure reconstruction performance, pure association discrepancy and our proposed association-based criterion respectively. *Fix* is to fix *Learnable* scale parameter σ of prior-association as 1.0. *Max* and *Minimax* ref to the strategies for association discrepancy in the maximization (Equation 4) and minimax (Equation 5) way respectively.

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**

(1) Anomaly Criterion **18.76%↑** (2) Prior-association **8.43%↑** (3) optimization strategy **7.84%↑**

Anomaly Criterion 可视化



Prior-Association 可视化

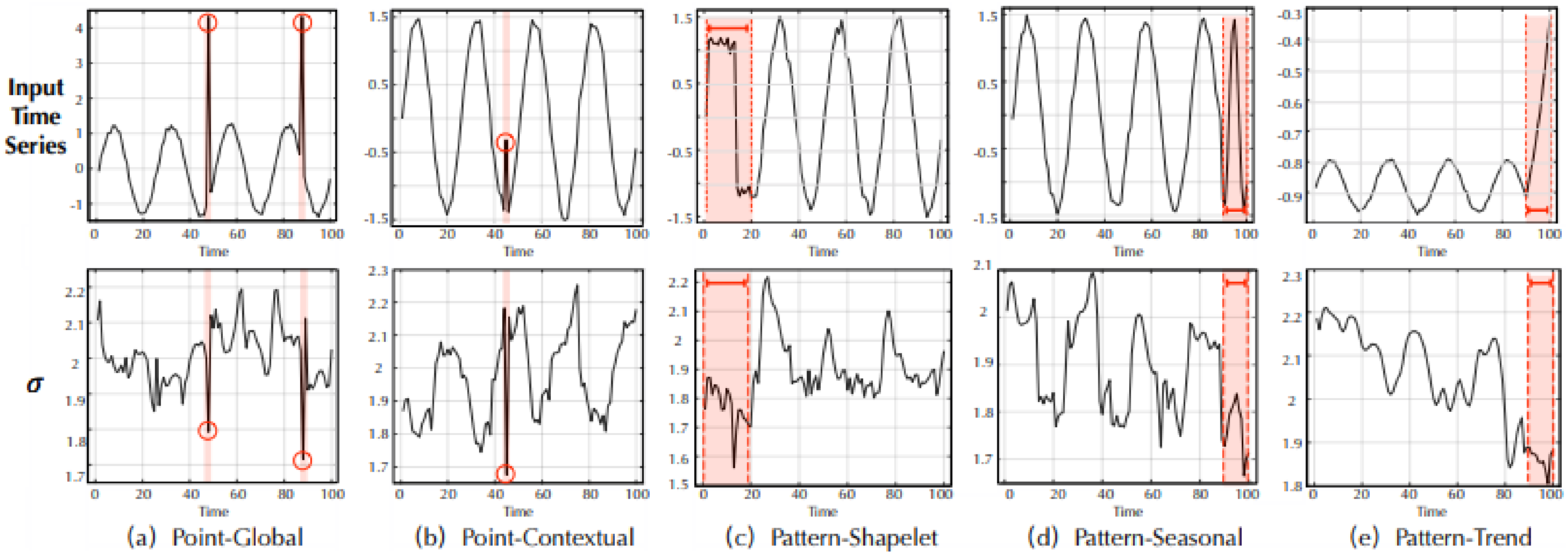


Figure 6: Learned scale parameter σ for different types of anomalies (highlight in red).

Prior-association can adapt to various data patterns of time series.

Abnormal time points show the adjacent-concentration property.

异常点的尺度参数 σ 是比较小的，意味着异常点更关注于临近部分。

Optimization Strategy的统计

Table 3: The statistical results of adjacent association weights for *Abnormal* and *Normal* time points respectively. *Recon*, *Max* and *Minimax* represent the association learning process that is supervised by reconstruction loss, direct maximization and minimax strategy respectively. A higher contrast value ($\frac{\text{Abnormal}}{\text{Normal}}$) indicates a stronger distinguishability between normal and abnormal time points.

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**

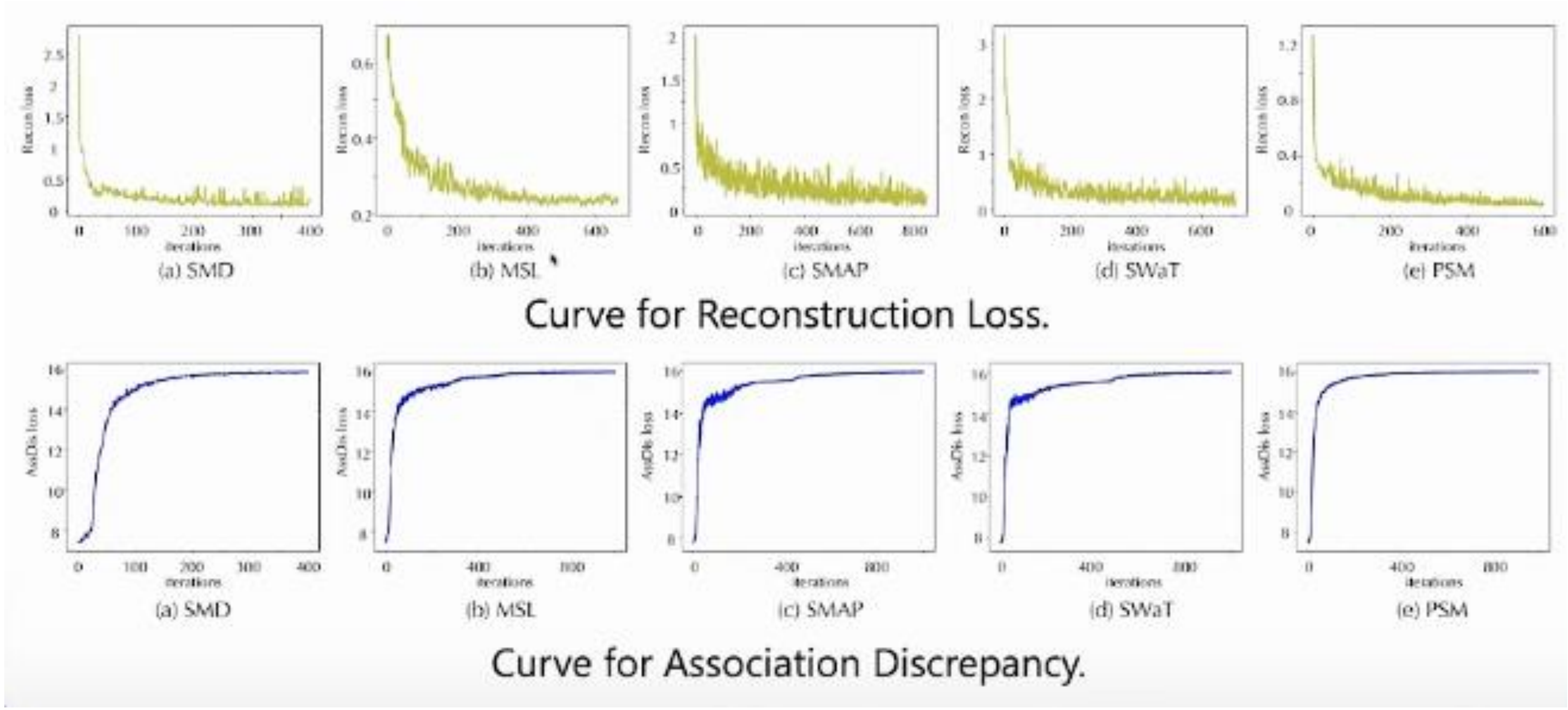
直接最大化关联差异会导致退化。
Minimax association 学习将放大normal-abnormal 的可区分性。

Optimization Strategy的统计

Table 3: The statistical results of adjacent association weights for *Abnormal* and *Normal* time points respectively. *Recon*, *Max* and *Minimax* represent the association learning process that is supervised by reconstruction loss, direct maximization and minimax strategy respectively. A higher contrast value ($\frac{\text{Abnormal}}{\text{Normal}}$) indicates a stronger distinguishability between normal and abnormal time points.

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**

Minimax 训练稳定性



关联差异的消融

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
L2	85.26	74.80	79.69	85.58	81.30	83.39	91.25	56.77	70.00	79.90	87.45	83.51	70.24	96.34	81.24
CE	88.23	81.85	84.92	90.07	86.44	88.22	92.37	64.08	75.67	62.78	81.50	70.93	70.71	94.68	80.96
Wasserstein	78.80	71.86	75.17	60.77	36.47	45.58	90.46	57.62	70.40	92.00	71.63	80.55	68.25	92.18	78.43
JSD	85.33	90.09	87.64	91.19	92.42	91.80	94.83	95.14	94.98	83.75	96.75	89.78	95.33	98.58	96.93
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**

Symmetrized KL divergence is the best choice.

研究了不同的距离函数

关联判据的消融

$$\text{Association Discrepancy: AnomalyScore}(\mathcal{X}) = \text{Softmax}\Big(-\text{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\Big),$$

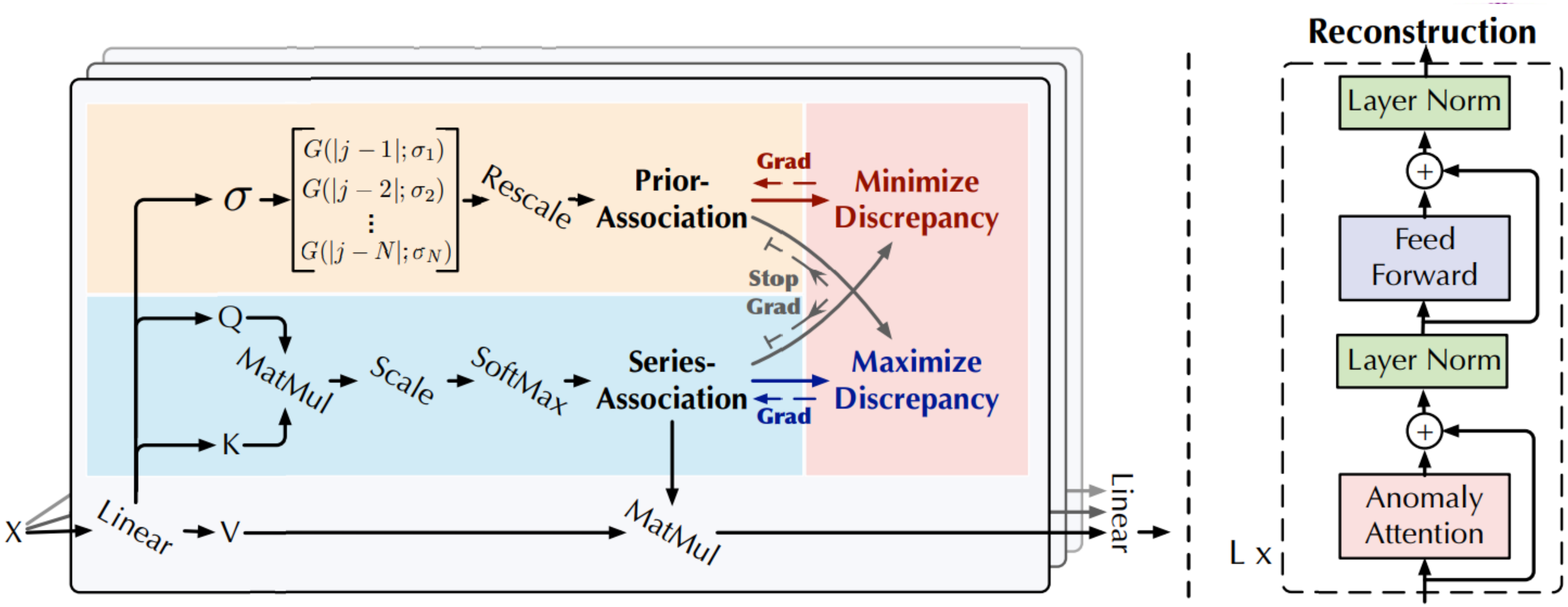
$$\text{Reconstruction: AnomalyScore}(\mathcal{X}) = \Big[\|\mathcal{X}_{i,:} - \widehat{\mathcal{X}}_{i,:}\|_2^2\Big]_{i=1,\cdots,N},$$

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

$$\text{Multiplication (Ours): AnomalyScore}(\mathcal{X}) = \text{Softmax}\Big(-\text{AssDis}(\mathcal{P},\mathcal{S};\mathcal{X})\Big)\odot\Big[\|\mathcal{X}_{i,:} - \widehat{\mathcal{X}}_{i,:}\|_2^2\Big]_{i=1,\cdots,N}. \tag{7}$$

Dataset	SMD			MSL			SMAP			SWaT			PSM			Avg
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1(%)
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	88.01
Recon	78.63	65.29	71.35	79.15	78.07	78.61	89.38	56.35	69.12	76.81	86.89	81.53	69.84	94.73	80.40	76.20
AssDis	86.74	88.42	87.57	91.20	89.81	90.50	91.56	90.41	90.98	97.27	89.48	93.21	97.80	93.25	95.47	91.55
Addition	77.16	70.58	73.73	88.08	87.37	87.72	91.28	55.97	69.39	84.34	81.98	83.14	97.60	97.61	97.61	82.32
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**	**94.96**

Anomaly Transformer 总结



(1)架构:具有 Anomaly-Attention机制 的Anomaly Transformer

(2)训练策略: Minimax 关联学习

(3)准则: 基于关联的异常准则

提供了一个全新的 association 视角来看待 时间序列异常检测

参考资料

1. **Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy**
2. ICLR 2022分享会-吴海旭-基于关联差异的时序异常检测算法
3. 作者讲解PPT
4. 论文十问

xjh20@mails.tsinghua.edu.cn
whx20@mails.tsinghua.edu.cn

长按关注，获取最新资讯