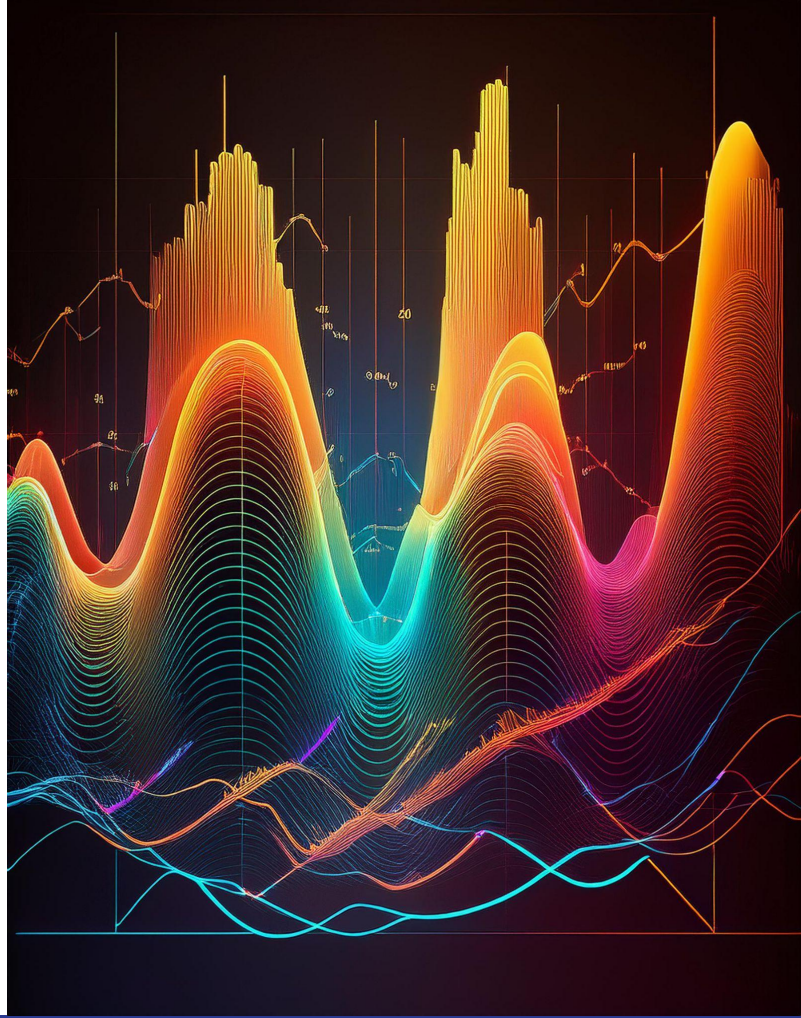
Anomaly Transformer
un'architettura basata sull'attenzione per il rilevamento di anomalie nelle serie temporali

Bruno Guzzo, Mat. 242504

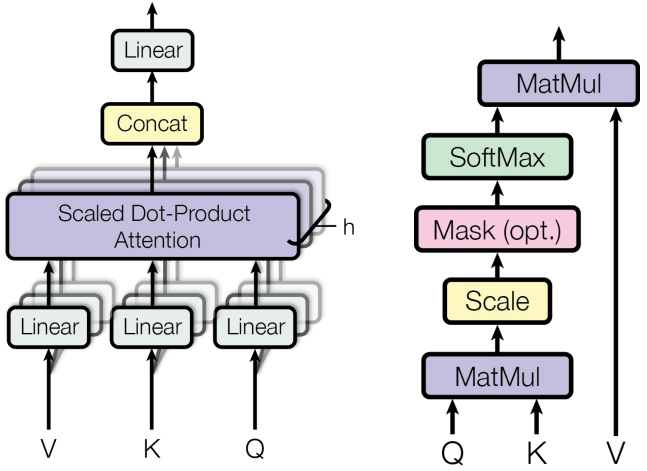
Introduzione

- Il rilevamento di anomalie nelle serie temporali è cruciale in molti settori.
- **Anomaly Transformer** è un modello innovativo basato sull'**attenzione** per affrontare questa sfida in modo non supervisionato.
- L'idea chiave è che le anomalie hanno difficoltà a stabilire **relazioni a lungo termine**, concentrandosi principalmente sui punti adiacenti (bias di concentrazione adiacente).



Il meccanismo dell'attenzione

- L'attenzione permette di **pesare dinamicamente** l'importanza di diverse parti dell'input.
- Cattura relazioni a lungo termine e modella dipendenze complesse.
- La **Scaled Dot-Product Attention** è una delle implementazioni più comuni.
- La **self-attention multi-head** consente di considerare diverse rappresentazioni dell'input.

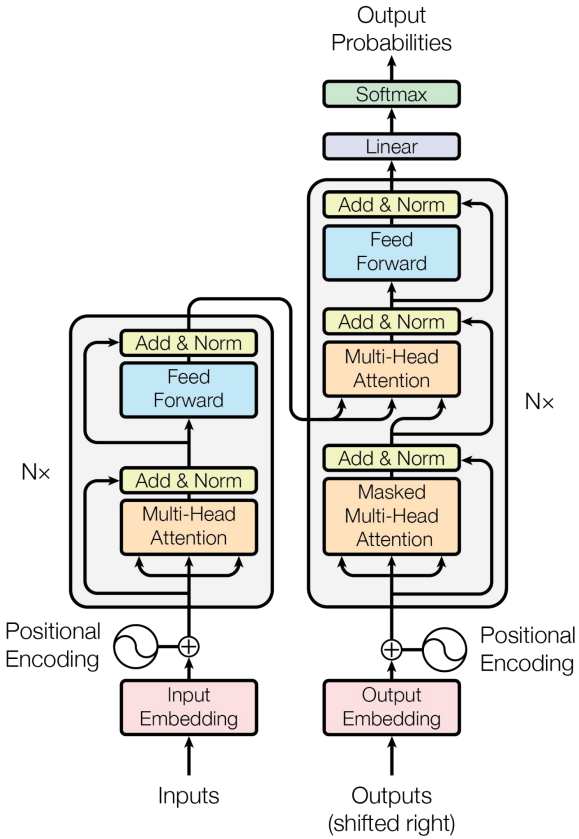


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Il Transformer

- Un nuovo modello di **rete neurale** basato interamente su meccanismi di **attenzione**, eliminando la necessità di **ricorrenza** e **convoluzioni**.
- Consente una **maggiore parallelizzazione**, accelerando l'addestramento e migliorando le prestazioni in attività di trasduzione di sequenze come la traduzione automatica.



Transformer - architettura del modello

Architettura dell'Anomaly Transformer

- Struttura simile ai modelli **Transformer**, ma senza decoder.
- Introduce l'**Anomaly Attention** per calcolare la discrepanza di associazione.
- Modella l'associazione a priori (**prior-association**) e l'associazione di serie (**series-association**).
- La **discrepanza** tra queste due associazioni è la base per il rilevamento delle anomalie.

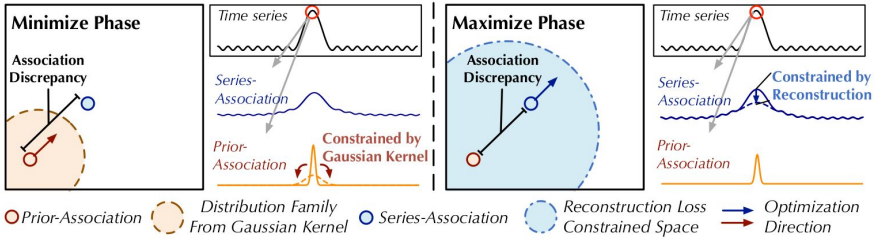
$$\text{Initialization: } \mathcal{Q}, \mathcal{K}, \mathcal{V}, \sigma = \mathcal{X}^{l-1}W_{\mathcal{Q}}^{l}, \mathcal{X}^{l-1}W_{\mathcal{K}}^{l}, \mathcal{X}^{l-1}W_{\mathcal{V}}^{l}, \mathcal{X}^{l-1}W_{\sigma}^{l}$$

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

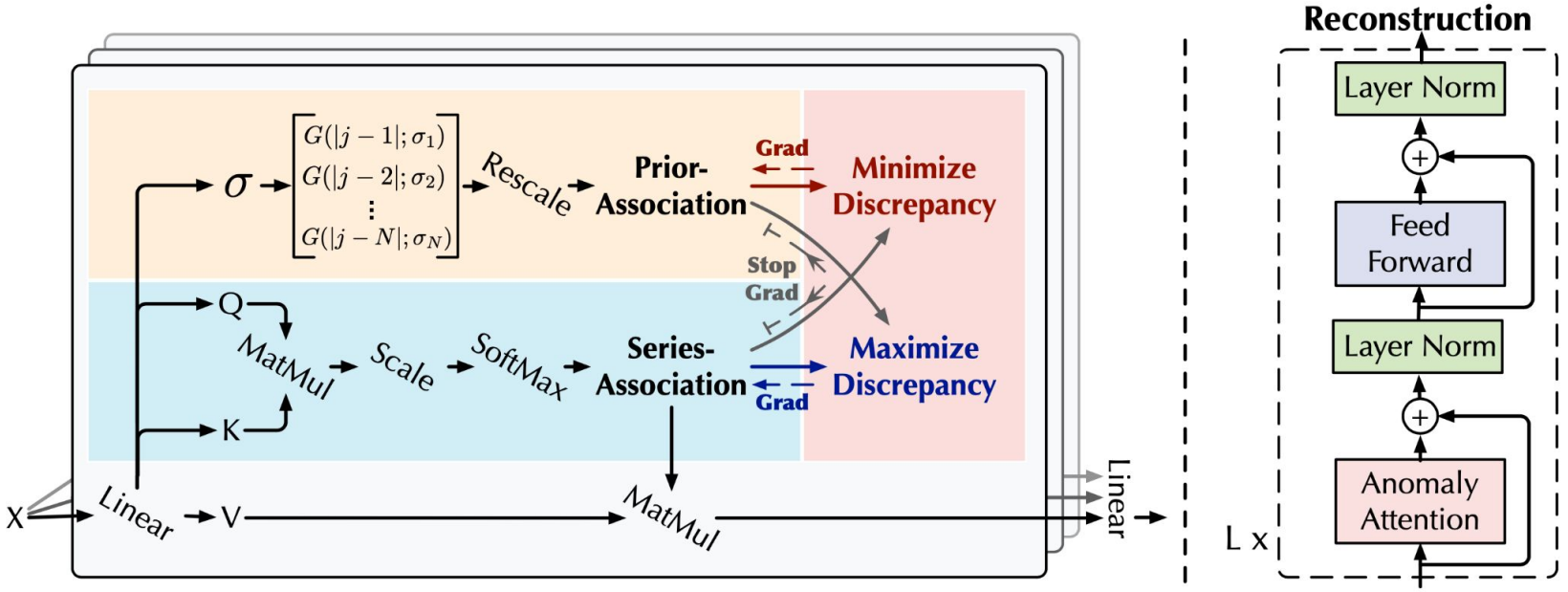
$$\text{Series-Association: } \mathcal{S}^{l} = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right)$$

$$\text{Reconstruction: } \widehat{\mathcal{Z}}^{l} = \mathcal{S}^{l}\mathcal{V},$$

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



Architettura dell'Anomaly Transformer



Apprendimento Minimax dell'Associazione

- Migliora la capacità di distinguere tra punti normali e anomali.
- **Fase di minimizzazione**: la *prior-association* approssima la *series-association*.
- **Fase di massimizzazione**: la *series-association* aumenta la *discrepanza di associazione*.
- La **funzione di perdita** bilancia l'errore di ricostruzione e la discrepanza di associazione.

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

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

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

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

Analisi di sensitività degli iperparametri

- Studio del comportamento del modello con **dimensionalità ridotta**.
- **Grid search** per testare diverse combinazioni di iperparametri.
- Non si rilevano miglioramenti significativi in termini di *F-Score*.
- Il modello si dimostra **robusto** a varie configurazioni, grazie alla strategia **minimax**.

Learning rate: 10^-4
Lambda : 0.5, 3
Anomaly ratio: 1
Epochs: 2
Batch Size: 128
Dimensione dmodel: 64
Numero di livelli l: 3
Numero di attention heads h: 8
Tipo di Kernel: Gaussiano, No (prior-association come parametro appreso) e sigmoide.
Funzione di perdita: Norma L1, Norma L2 (MSE), Cross Entropy e Divergenza KL

	Accuracy	Precision	Recall	F-Score
$d_{model} = 512$ $epoch = 3$	0.9845047	0.9181818	0.9363893	**0.9271962**
$d_{model} = 64$ $epoch = 2$	0.9865807	0.9205660	0.9550605	**0.9374961**

Algoritmi di ottimizzazione

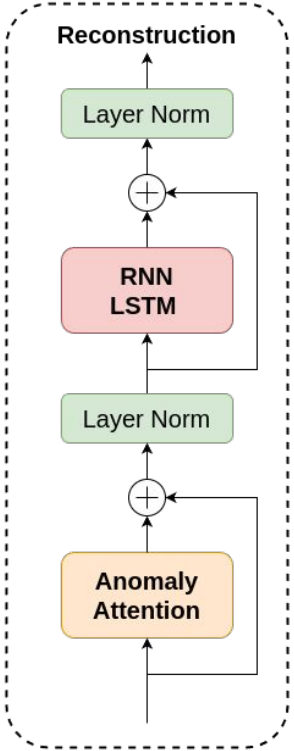
- **ADAM** è l'algoritmo scelto dagli autori, efficiente per problemi di grandi dimensioni.
- Confronto con altri algoritmi: *SGD*, *RMSprop*, *Adadelta*, *AdamW*.
- ADAM si conferma il più performante in termini di **tempo di addestramento**.

	Accuracy	F-Score	Train Time (S)
Adam	0.9863501	0.9364257	**199.2156**
AdamW	0.9862144	0.9357856	**199.3804**
SGD	0.9896608	0.9525943	**199.6484**
Adadelta	0.9875848	0.9425360	**200.2806**
RMSprop	0.9857802	0.9336793	**200.7332**

Uso di RNN in combinazione con Anomaly Transformer

- Introduzione di reti **LSTM** per cercare relazioni a lungo/breve termine.
- **Nessun miglioramento** sostanziale rispetto al modello originale.
- La strategia **minimax** si rivela efficace anche con modifiche infrastrutturali.

RNN Layers	Accuracy	F-Score	Train Time (S)
0	0.9874491	**0.9418714**	199.9097
1	0.9892809	**0.9508217**	202.9323
2	0.9875441	**0.9425028**	208.6800
4	0.9880054	**0.9445211**	216.3183
8	0.9849254	**0.9292131**	234.4055
16	0.9871642	**0.9405256**	271.9105

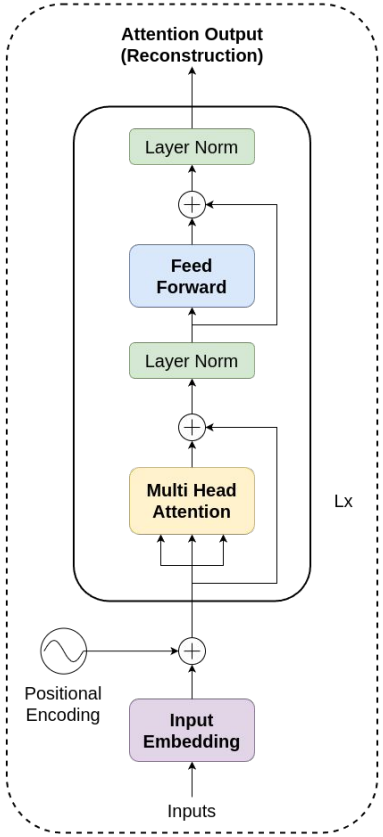


Anomaly Attention vs Self Attention

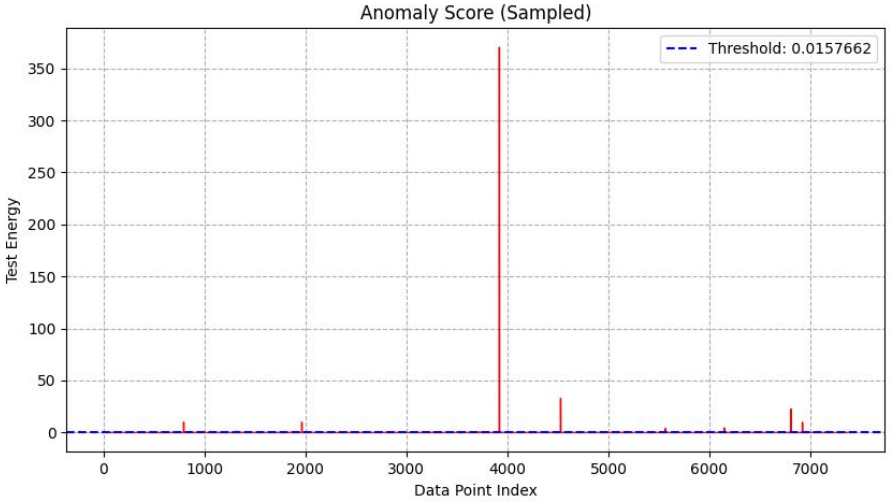
- Confronto tra **Anomaly Attention** e **Self Attention** classica.
- Il modello con Anomaly Attention offre **prestazioni superiori** del 43%!
- L'**Anomaly Score** stabilizza l'errore di ricostruzione, facilitando l'identificazione delle anomalie.

$$AnomalyScore(X) = \left[\left\| X_{i,:} - \widehat{X}_{i,:} \right\|_2^2 \right]_{i=1,\ldots,N}$$

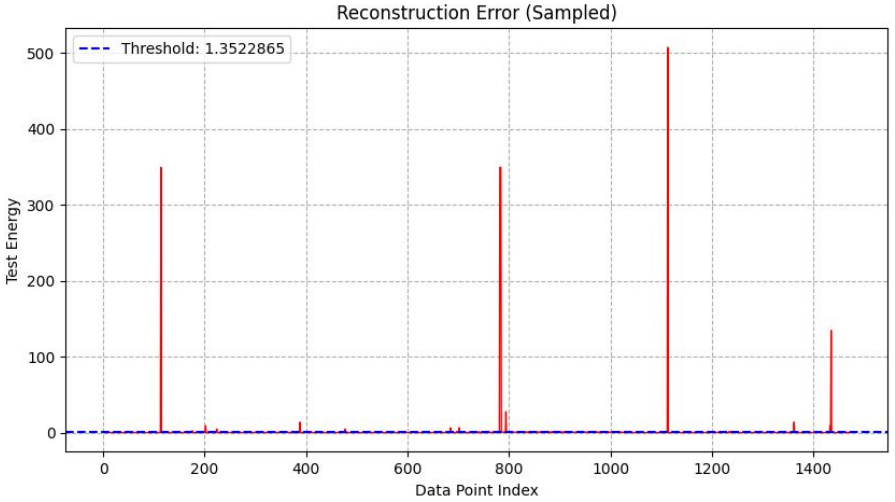
Modello	Accuracy	F-Score	Train Time
Self-attention $d_{model} = 128$ $epoch = 3$	0.8515604	**0.5525562**	38.5978
Self-attention $d_{model} = 512$ $epoch = 5$	0.7999457	**0.4547337**	352.9563
Anomaly Attention $d_{model} = 128$ $epoch = 3$	0.9883039	**0.9460643**	312.0443
Anomaly Attention $d_{model} = 512$ $epoch = 5$	0.9840706	**0.9249648**	645.7070



Anomaly Attention vs Self Attention



Anomaly Transformer



Modello che fa uso di **self-attention** classica

Conclusioni

- L'Anomaly Transformer è un **avanzamento significativo** nel rilevamento di anomalie.
- Cattura relazioni complesse e **distingue efficacemente** le anomalie.
- La **strategia minimax** e l'**Anomaly Attention** sono i **punti di forza** del modello.
- Apre **nuove prospettive** per l'applicazione del deep learning in contesti reali.