FormerTime: Hierarchical Multi-Scale Representations for Multivariate Time Series Classification

Reporter: Mingyue Cheng

Mingyue Cheng, Qi Liu*, Zhiding Liu, Zhi Li, Yucong Luo, Enhong Chen

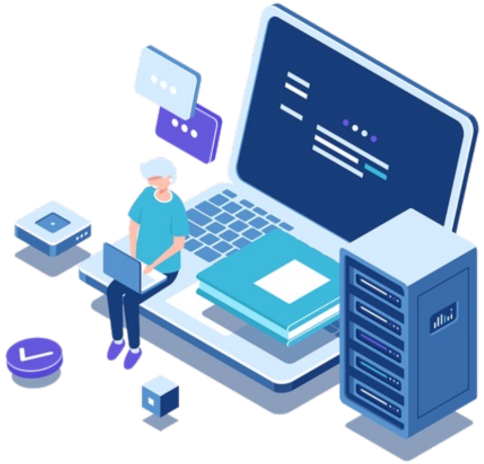
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence, Hefei, China

Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

Background

□ Time series

- Time series data are sequences of observations collected over time, have been the subject of significant research interest in recent years due to their importance in various domains.
- The analysis of time series data not only has significant academic research value but also is an essential tool for data-driven decision-making broad range of applications.



Abnormal Traffic Detection

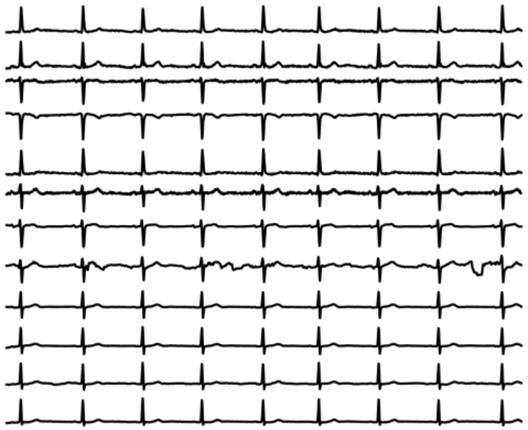
Healthcare Monitoring

Industrial Detection

Problem Definition

- **Time series classification (TSC)**
  - The goal of TSC is to build a function model that can learn the patterns in the time series data and generalize well to make accurate predictions on unseen data.
- **Find informative patterns relative to target class labels**
  - It usually refers to various complex patterns to be mined for TSC
  - A typical feature is to cover different time scales
    - ✓ Local sub-series or long time interval

Sinus Rhythm



Sinus Bradycardia

Example of ECG Classification

Related Works & Motivation

- Traditional methods
  - Shapelet based model
    - Learning shapelets
  - Distance-based
    - NN-DTW
  - Feature based
    - XGBoost

- Key limitations
  - ✓ Expensive computation cost
  - ✓ Hard to serve large scale time series scenario
  - ✓ Linear transformation

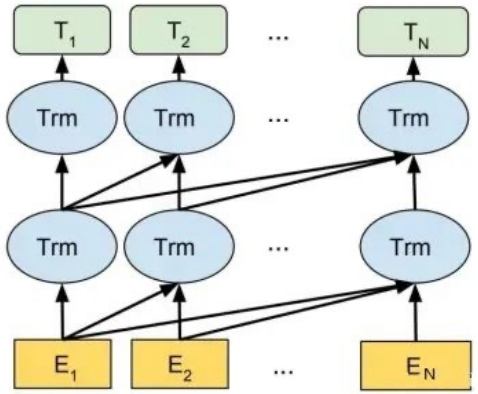
- Convolution neural networks (CNNs) play a vital role in time series classification 😊
  - Three aspects of strength w.r.t. applying CNNs in time series classification
    - ✓ Multi-scale representations with varying strides
    - ✓ Weight-sharing mechanism
    - ✓ Can be computed in parallel 😰
- One key limitation
  - <u>Lacking of the capacity of global context modeling</u>

- Key strengths of deep learning models
  - ✓ Can easily scale to large-scale data
  - ✓ Non-linear transformation capacity

Ruiz, Alejandro Pasos, et al. "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances." *Data Mining and Knowledge Discovery* 35.2 (2021): 401-449.

Related Works & Motivation

- Transformer models
  - Transformer models preserve the strong capacity of global contexts and has achieved great success in language text representation.
- Challenges in adapting Transformers from language to time series
  - Basic semantic unit: **human-generated discrete word v.s. temporal continuous value.**
  - Sequence length: **very short or limited sequence length v.s. very long sequences.**
  - Position information: **only sequence v.s. time property.**
- Drawback of TST [Zerveas et al, KDD2021], which a transformer-based framework proposed for time series classification
  - Expensive computation cost
    - Its computation cost is sequence length
  - Lack of multi-scale representations
    - Lack of hierarchical architecture
  - Weak translation invariance capacity
    - Dynamic weight instead of weight-sharing



Zerveas, George, et al. "A transformer-based framework for multivariate time series representation learning." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.

Overview of the FormerTime

Temporal slicing partition operation and multi-scale representation.

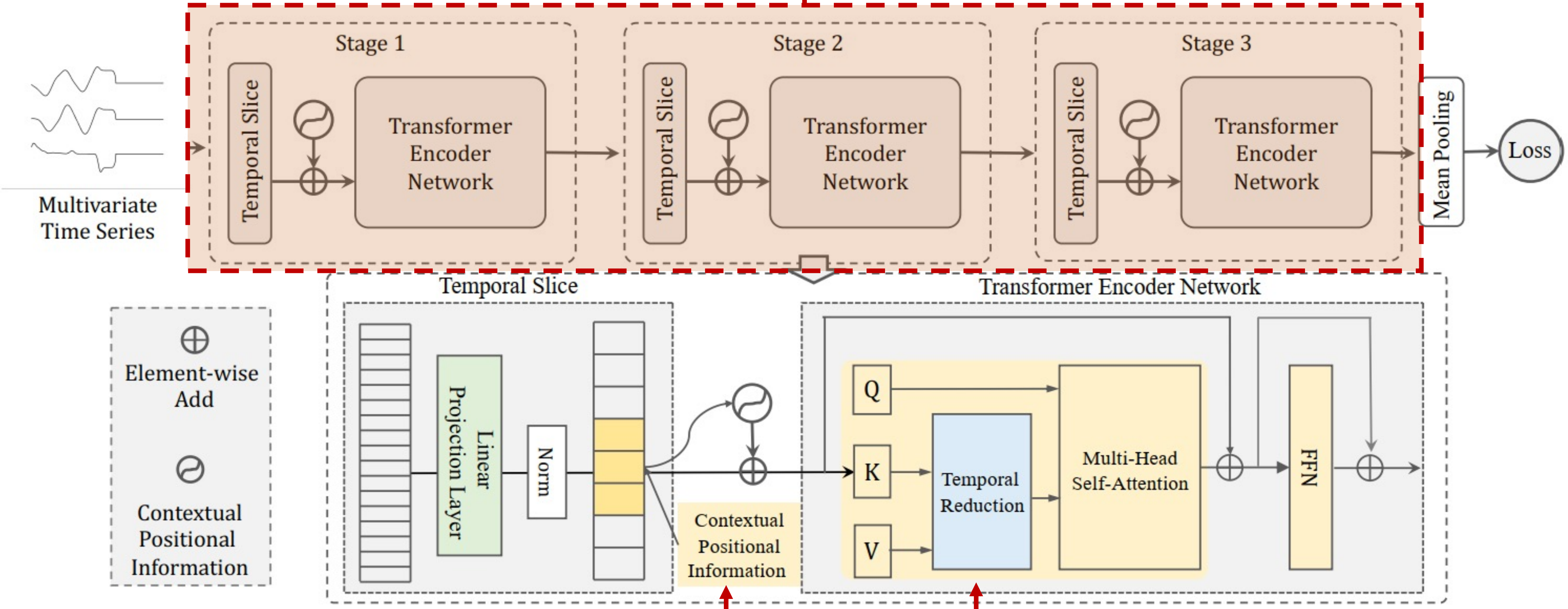


Figure 1: Illustration of the FormerTime, i.e., a efficient hierarchical transformer architecture for the MTSC task.

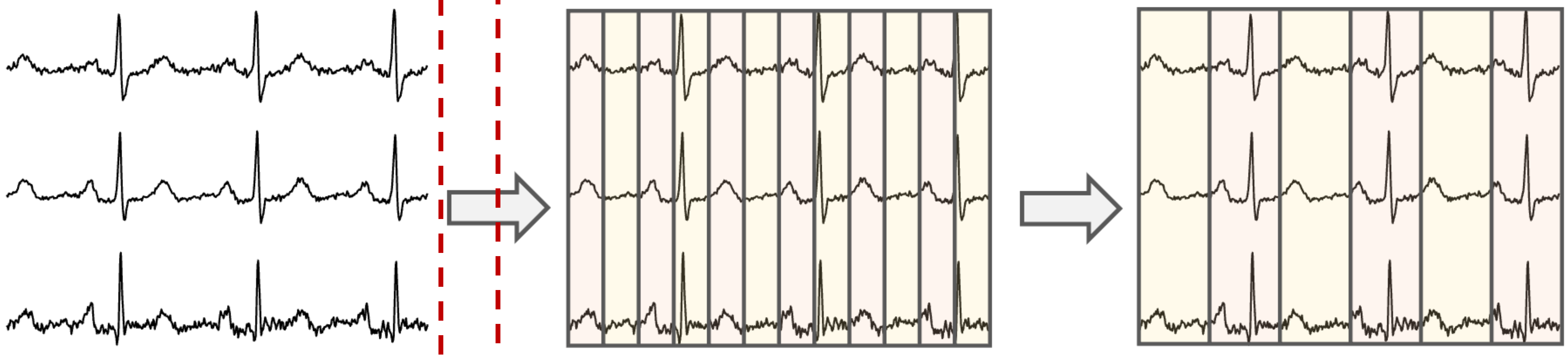
Contextual positional encoding strategies.

Temporal reduction attention layer.

Temporal Slice Partition

Temporal slicing partition: time series point in local regions are modeled together instead of individually learning their representation.

Stage-wise network architecture: the varying scale of time series data can be effectively learned by flexibly updating the number of stages.



Raw time series

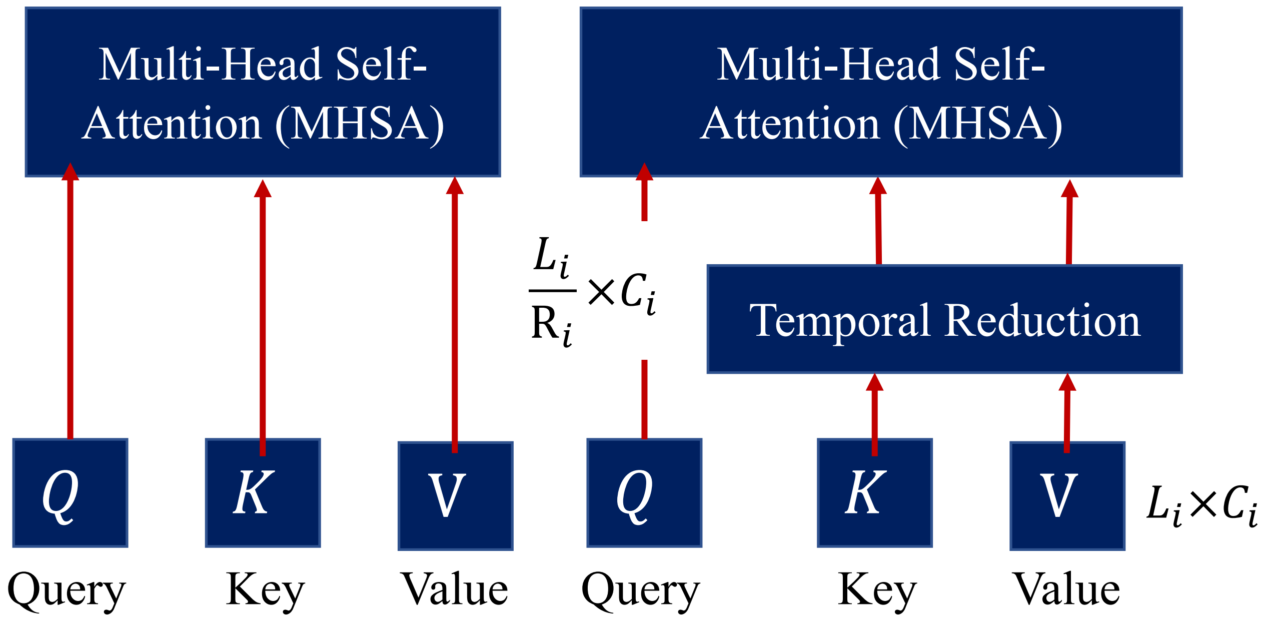
Temporal slice partition with fine-grained

Temporal slice partition with rough scale

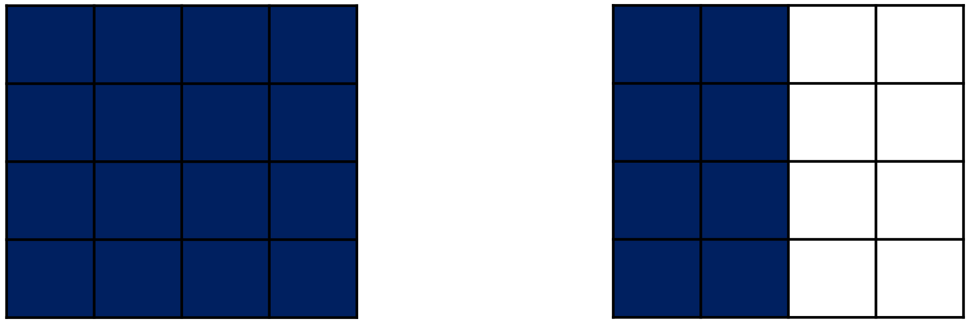
Single scale representations.
Expensive computation costs.

Multi-scale representation transformation.
The sequence length is largely reduced.

A Novel Transformer Encoder



Vanilla Multi-head Self-Attention Layer ⇨ Temporal Reduction Multi-head Self-Attention Layer



- ✓ Making the input sequence permutation-variant but temporal invariant is a necessity for time series classification.
- ✓ Having the ability to provide absolute information also matters.

Experiment Settings

- Datasets
  - Ten public time series classification datasets chosen from UEA archive.
- Evaluation Metrics
  - Classification performance
    - ✓ Accuracy
  - Computation cost
    - ✓ MACs
- Compared Baselines
  - Shapelet-based methods
    - ✓ Learning Shapelets
    - ✓ Shapelet Transformation
  - Convolution-based methods
    - ✓ MDCNN
    - ✓ InceptionTime
    - ✓ MiniRocket
  - Self-attention based methods
    - ✓ TST/Informer/GTN

Table 1: Statics of datasets in the experiments.

Dataset	Train Size	Test Size	Dimensions	Length	Classes
AWR	275	300	9	144	25
AF	15	15	2	640	3
CT	1,422	1,436	3	182	20
CR	108	72	6	1,197	12
FD	5,890	3,524	144	62	2
FM	316	100	28	50	2
MI	278	100	64	3,000	2
SRS1	268	293	6	896	2
SRS2	200	180	7	1,152	2
UWG	120	320	3	315	8

Experimental Results

Table 3: Classification performance of compared methods in ten datasets. Bold numbers represent the best results.

Datasets	IT	LS	ST	MCDCNN	TCN	MCNN	ResNet	MR	TST	GTN	Informer	Ours
AWR	0.9827	0.9127	0.8700	0.7800	0.9467	0.8200	0.9827	0.9720	0.9789	0.9767	0.9820	**0.9847**
AF	0.4400	0.2533	0.2667	0.3733	0.4933	0.3467	0.4000	0.3333	0.4000	0.4000	0.4267	**0.6000**
CT	**0.9983**	0.9866	0.7224	0.8826	0.9915	0.9238	0.9965	0.9876	0.9882	0.9783	0.9862	0.9914
CR	0.9889	0.9639	0.9722	0.6278	0.9083	0.9167	**0.9972**	0.9806	0.9583	0.7917	0.9778	0.9806
FD	0.6820	0.5129	0.5085	0.5000	0.6801	0.6747	0.5760	0.6065	0.6005	0.5542	0.5265	**0.6872**
FM	0.6000	0.4840	0.4940	0.5920	0.5880	0.5920	0.6080	**0.6380**	0.5900	0.5350	0.6120	0.6180
MI	0.5860	0.5180	0.6100	0.5000	0.6040	0.5980	0.5780	0.5640	N/A	N/A	0.6240	**0.6320**
SRS1	0.8942	0.7038	0.6724	0.9079	0.9031	0.8949	0.8730	**0.9352**	0.8771	0.8019	0.9188	0.8867
SRS2	0.5689	0.5111	0.5300	0.5256	0.5978	**0.5989**	0.5622	0.5411	0.5796	0.5611	0.5767	0.5922
UWG	0.8869	0.8031	0.7769	0.8438	0.7981	0.8044	0.7994	**0.9075**	0.8271	0.8406	0.8363	0.8881
Average	0.7628	0.6649	0.6423	0.6533	0.7511	0.7170	0.7373	0.7466	0.7555	0.7155	0.7467	**0.7861**
MACs (M)	89	-	-	263	283	929	132	-	408	1,565	141	98

Our FormerTime can achieve superior classification accuracy in average, reflecting the potential application of Transformers in time series classification tasks.
The computation cost of FormerTime is only similar with convolutional based models.

Experimental Results

Studying the impact of stage number.

Table 4: Experimental results w.r.t. studying the hyper-parameter sensitivity with varying stages.

Datasets	1	2	3	4
AWR	**0.9811**	**0.9811**	0.9720	0.9767
AF	0.4222	0.4667	**0.6000**	0.5778
CT	0.9907	0.9909	**0.9914**	0.9902
CR	**0.9861**	0.9815	0.9806	0.9769
FD	0.6750	**0.6793**	0.6776	0.6748
FM	**0.6200**	0.6033	0.6140	0.6067
MI	0.6200	0.6267	**0.6280**	0.6133
SRS1	0.8760	0.8692	0.8771	**0.8840**
SRS2	0.5722	0.5815	**0.5922**	0.5889
UWG	**0.9021**	0.8948	0.8844	0.8844
Averge	0.7645	0.7675	**0.7817**	0.7774

- ✓ In terms of the hierarchical architecture, it seems to require different number of stage with respect to the specific datasets.
- ✓ In the UEA datasets, it can help us achieve superior performance while preserving the number of stage as 3.

Studying the impact of temporal slice size.

Table 5: Experimental results w.r.t. studying the hyper-parameter sensitivity w.r.t. temporal slice size.

Datasets	[16,32,64]	[8,16,32]	[4,8,16]	[2,4,8]
AWR	0.9720	0.9740	0.9820	**0.9847**
AF	**0.6000**	0.5600	0.4267	0.4400
CT	**0.9914**	0.9886	0.9868	0.9873
CR	**0.9806**	0.9806	0.9778	0.9667
FD	0.6776	**0.6794**	0.6823	0.6872
FM	0.6140	0.6080	**0.6180**	0.6040
MI	**0.6280**	**0.6280**	0.6160	0.6180
SRS1	0.8771	0.8826	0.8710	**0.8867**
SRS2	**0.5922**	0.5811	0.5856	0.5600
UWG	0.8844	**0.8881**	0.8781	0.8775
Averge	**0.7817**	0.7770	0.7624	0.7612

- ✓ It seems that larger slice size can help us achieve superior classification performance in most situations.
- ✓ This is most probably because larger slice size can further enhance the information density of sub-series.

Studying the effectiveness of our positional encodings.

Table 6: Experimental results w.r.t. studying the effectiveness of contextual positional embeddings.

Datasets	None	Static	Learnable	Ours
AWR	0.9433	0.9822	0.9811	**0.9720**
AF	0.4667	0.5111	0.5556	**0.6000**
CT	0.9821	0.9902	0.9863	**0.9914**
CR	**0.9815**	0.9676	0.9769	0.9806
FD	0.6740	**0.6804**	0.6774	0.6776
FM	0.5900	0.5867	**0.6200**	0.6140
MI	0.6233	0.5833	0.6167	**0.6280**
SRS1	0.8635	**0.8817**	0.8749	0.8771
SRS2	0.5704	0.5759	**0.6018**	0.5922
UWG	0.8479	0.8729	0.8677	**0.8844**
Averge	0.7543	0.7632	0.7758	**0.7817**

- ✓ Our contextual encoding strategy can exhibit other several prevalent methods of positional encoding methods.
- ✓ It indicates the essence of absolute and relevance of positional encoding strategy.

Experimental Results

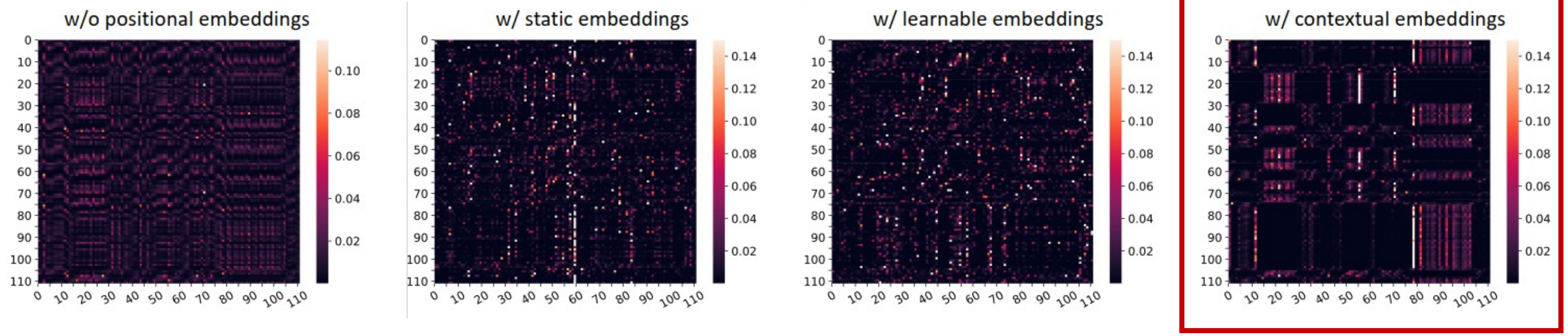


Figure 3: Normalized attention score from the first encoder block of the first stage in FormerTime: (1) without taking positional information into account, (2) using static embeddings, (3) using learnable vectors, (4) using our contextual embeddings.

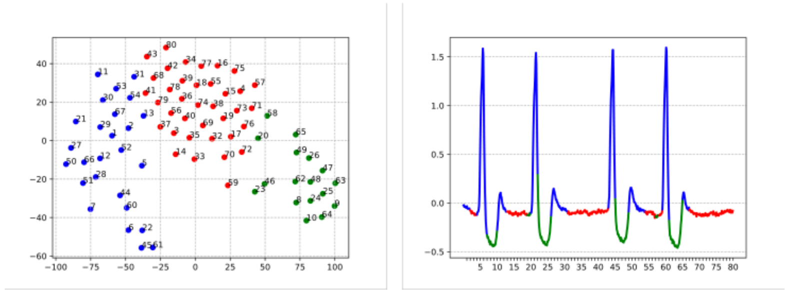


Figure 4: Left plot: Visualization of the t-SNE result of the embedding layer output on the AF dataset. Right plot: visualization of sub-sequences on raw time series data.

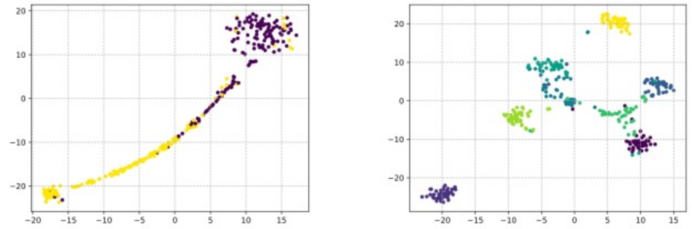


Figure 5: Visualization of the representation of whole time series on the SRS1 (left plot) and UW (right plot) datasets, extracted by pooling operation from the last hidden layer.

FormerTime can effectively capture the semantic information of sub-series.

FormerTime can learn high-quality representations of time series data via supervised learning.

Conclusion & Take Away Message

- We try to show the potential of applying Transformer network in the classification of time series so as to promote the development of time series mining.

- We proposed a novel Transformer based model for time series classification
  - Multi-scale representation of time series
    - ✓ Temporal slicing partition
    - ✓ Hierarchical network architecture
  - A novel Transformer encoder network
    - ✓ Contextual positional encoding
    - ✓ Temporal reduction attention layer

GitHub

https://github.com/Mingyue-Cheng/FormerTime

- We conduct extensive experiments on 10 UEA datasets
  - FormerTime can achieve superior performance for the classification of time series in average.
  - FormerTime can overcome the inefficient computation issue incurred by the original setting of feeding raw time series into vanilla self-attention mechanism.

Our Research Plan for Time Series Classification

Transformer-based Classification Network → **Self-supervised based Pre-trained Model** → **Multimodal based Universal Time Series Model**

FormerTime (WWW, 2023)

TimeMAE (Preprint, 2023)

Towards Universal Time Series Modeling (**Working**)

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO. 8, AUGUST 2022 1

TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders

Mingyue Cheng, Qi Liu*, Zhiding Liu, Hao Zhang, Rujiao Zhang, Enhong Chen

https://github.com/Mingyue-Cheng/TimeMAE

Cheng, Mingyue, et al. "TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders." *arXiv preprint arXiv:2303.00320* (2023).

References

- Zheng, Y., Liu, Q., Chen, E., Ge, Y. and Zhao, J.L., 2014. Time series classification using multi-channels deep convolutional neural networks. In Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15 (pp. 298-310). Springer International Publishing.
- Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A. and Petitjean, F., 2020. Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery, 34(6), pp.1936-1962.
- Dempster, A., Petitjean, F. and Webb, G.I., 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery, 34(5), pp.1454-1495.
- Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A. and Eickhoff, C., 2021, August. A transformer-based framework for multivariate time series representation learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 2114-2124).
- Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J. and Long, M., 2022. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. arXiv preprint arXiv:2210.02186.
- Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M. and Bagnall, A., 2021. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 35(2), pp.401-449.
- Bagnall, A., Lines, J., Bostrom, A., Large, J. and Keogh, E., 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data mining and knowledge discovery, 31, pp.606-660.
- He, W., Cheng, M., Liu, Q. and Li, Z., 2023, April. ShapeWordNet: An Interpretable Shapelet Neural Network for Physiological Signal Classification. In Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part IV (pp. 353-369). Cham: Springer Nature Switzerland.
- Cheng, M., Liu, Q., Liu, Z., Li, Z., Luo, Y. and Chen, E., 2023. FormerTime: Hierarchical Multi-Scale Representations for Multivariate Time Series Classification. arXiv preprint arXiv:2302.09818.
- Cheng, M., Liu, Q., Liu, Z., Zhang, H., Zhang, R. and Chen, E., 2023. TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders. arXiv preprint arXiv:2303.00320.

Thank You for Your Attention
Q&A

Research Homepage: https://mingyue-cheng.github.io/

Email: mycheng@mail.ustc.edu.cn