Transformers for Modeling Long-Term Dependencies in Time Series Data: A Review

S. Thundiyil[1], S.S. Shalamzari[2], J. Picone[2] and S. McKenzie[3]

1. Dept. of Elect. and Comm. Eng., BMS Institute of Technology and Management, Bengaluru, India
2. The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
3. University of New Mexico Health Sciences, Albuquerque, New Mexico, USA

saneesh@bmsit.in, {somayeh.seifi.shalamzari, picone}@temple.edu, sammckenzie@salud.unm.edu

Analysis of time series data for classification or prediction tasks is very useful in various applications such as healthcare, climate studies and finance. As big data resources have recently become available in a number of fields such as healthcare [1], finance [3]-[5] and climate change [6], it is now possible to apply state of the art deep learning models. Traditional methods such as autoregressive integrated moving average (ARIMA) [7], long short-term memory networks (LSTM) [8], gated recurrent units (GRUs) [9] and recurrent neural networks (RNN) [10] have provided robust frameworks in the analysis of time series data. However, these methods have limitations when applied to big data sets and when used to model long-term dependencies. The emergence of transformer-based architectures [11], as show in Figure 1, and technologies such as ChatGPT [12], has demonstrated the potential for analyzing time series data with long-term dependencies and advancing the basic science by discovering new underlying structure. In this review, we provide a detailed analysis of state of the art in deep learning systems that model long-term context.

Time series analysis techniques are evolving rapidly. Historically popular approaches such as LSTMs and GRUs suffer from the vanishing gradient problem when attempting to model extremely long-term phenomena. Systems based on an attention mechanism [11] leverage positional embedding modules and have been effectively employed in raw EEG data classification related to motor imagery tasks. A transformer-based architecture using a multi-head self-attention mechanism has been shown to provide promising levels of accuracy [13]. Furthermore, a novel decoding method called Spatial-Temporal Tiny Transformer (S3T), has highlighted the use of attention mechanisms [14].

Similarly, the combination of a self-supervised learning task and transformer models appear to be promising [15]. Transformer networks have been implemented to improve the performance and explainability of automatic seizure detection models, especially for continuous, long-term intracranial electroencephalogram (iEEG) data [16]. Impressive results, with high event-based sensitivity and low false positive rates, were demonstrated across two iEEG datasets. Consequently, the benefits of deep learning and transformer models are being recognized in commercial settings. A comparative analysis of commercial seizure-detection software packages like Besa 2.0, Encevis 1.7, and Persyst 13 revealed no significant difference in their per-patient detection rates [13]. However, this study pointed out significant variance in the false alarm rate, underlining the need for continued improvement in commercial offerings.

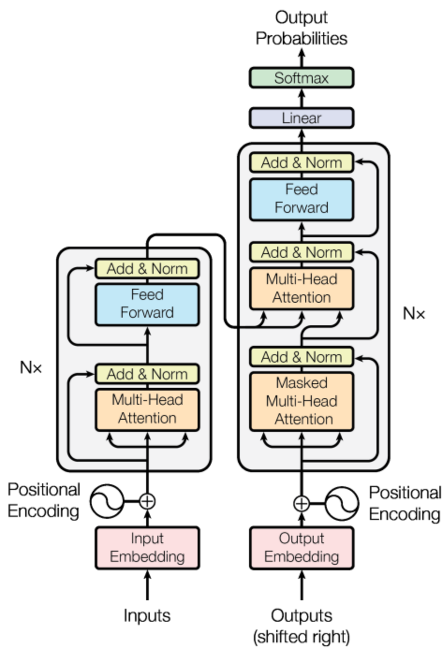


Figure 1. The original transformer model proposed in [11]

The ability of large memory models to capture long-term dependencies in time-series data enables the detection of subtle, complex patterns that may be indicative of impending seizure activity. This capability is particularly crucial in detecting seizures, which may be overlooked by traditional methods. Furthermore, these models can be trained on large, diverse datasets, allowing them to generalize across a wide range of seizure types and patient populations. Figure 2 provides a summary of the various application areas where transformer-based architectures are used, along with the most popular architectures for each application. Different applications in signal processing that addresses time-series related tasks such as forecasting, classification, and anomaly detection will be compared in the final review paper. The review will also compare popular variations of transformers including Vanila [17], LogTrans [18], InParformer [19], Informer [20], Sageformer [21], Autoformer [22], Pyraformer [23], W-Transformers [24], Quatformer [25], FEDformer [26], and Crossformer [27].

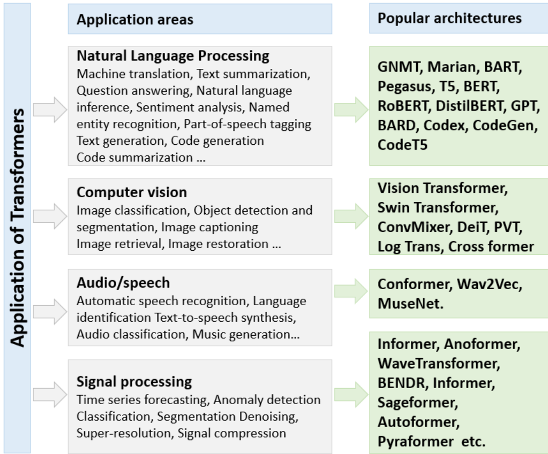


Figure 2. Popular transformer architectures and application areas

The success of the attention mechanism in natural language processing has motivated researchers to apply this technique to many fields involving time series analysis. Even though the attention mechanism has demonstrated the ability to capture temporal patterns over long periods of time, research in this area is relatively nascent and evolving rapidly. Systems based on these architectures are extremely complex and difficult to implement. As the data resources available for training are growing exponentially, the optimization of these architectures for specific applications will continue to pose challenges. The goal of this review is to make this new generation of technology more accessible to the community. Updates to this review will be available on GitHub at https://github.com/sanect/timeSeries_transformers.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation under Grant No. 2211841. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] J. Picone, S. McKenzie, and M. Desai, "Unencumbered Electroencephalogram (EEG) Corpora: Advancing Technology Using Big Data Resources," in International Neuroinformatics Coordinating Facility Neuromatics Conference (INCF), Stockholm, Sweden, 2022, p. 11. url: *https://isip.piconepress.com/publications/presentations_invited/2022/incf/tuh_eeg/*.

[2] A. Harati, S. Lopez, I. Obeid, J. Picone, M. P. Jacobson, and S. Tobochnik, "The TUH EEG CORPUS: A big data resource for automated EEG interpretation," in 2014 IEEE Signal Processing in Medicine and Biology Symposium, Dec. 2015, pp. 1–5. doi: *10.1109/SPMB.2014.7002953*.

[3] J. Fulcher, M. Zhang, and S. Xu, "Application of Higher-Order Neural Networks to Financial Time-Series Prediction," in *Artificial Neural Networks in Finance and Manufacturing*, J. Kamruzzaman, R. Begg, and R. Sarker, Eds., Hershey, PA, USA: IGI Global, 2006, pp. 80–108. doi: *10.4018/978-1-59140-670-9.ch005*.

[4] A. H. Bukhari, M. A. Z. Raja, M. Sulaiman, S. Islam, M. Shoaib, and P. Kumam, "Fractional Neuro-Sequential ARFIMA-LSTM for Financial Market Forecasting," *IEEE Access*, vol. 8, pp. 71326–71338, 2020. doi: *10.1109/ACCESS.2020.2985763*.

[5] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, p. 3, Jan. 2018. doi: *10.1186/s40537-017-0111-6*.

[6] V. Sebestyén, T. Czvetkó, and J. Abonyi, "The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking," *Frontiers in Environmental Science*, vol. 9, 2021. doi: *10.3389/fenvs.2021.619092*.

[7] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Melbourne, Australia: Otexts, 2021. url: *https://otexts.com/fpp3/*.

[8] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Proceedings of the Annual Conference of the International Speech Communication Association* (INTERSPEECH), Singapore, 2014, pp. 338-342. doi: *https://www.isca-speech.org/archive/interspeech_2014/i14_0338.html*.

[9] Y. Gao and D. Glowacka, "Deep Gate Recurrent Neural Network," in *Proceedings of the 8th Asian Conference on Machine Learning*, R. J. Durrant and K.E. Kim, Eds., PMLR, Nov. 2016, pp. 350-365. [Online]. url: *https://proceedings.mlr.press/v63/gao30.html*.

[10] L. Vidyaratne, A. Glandon, M. Alam, and K. M. Iftekharuddin, "Deep recurrent neural network for seizure detection," in *Proceedings of the International Joint Conference on Neural Networks* (IJCNN), Vancouver, British Columbia, Canada, 2016, pp. 1202–1207. doi: *10.1109/IJCNN.2016.7727334*.

[11] A. Vaswani et al., "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17). Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. doi: 10.5555/3295222.3295349.

[12] OpenAI, "GPT-4 Technical Report." 2023. [Online]. url: https://arxiv.org/abs/2303.08774

[13] J. Koren, S. Hafner, M. Feigl, and C. Baumgartner, "Systematic analysis and comparison of commercial seizure-detection software," *Epilepsia*, vol. 62, no. 2, 2021, doi: *10.1111/epi.16812*.

[14] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based Spatial-Temporal Feature Learning for EEG Decoding" *arXiv*, Jun. 10, 2021. doi: *10.48550/arXiv.2106.11170*.

[15] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data," *Front. Hum. Neurosci.*, vol. 15, Jun. 2021, doi: *10.3389/fnhum.2021.653659*.

[16] Y. Sun et al., "Continuous Seizure Detection Based on Transformer and Long-Term iEEG," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5418–5427, Nov. 2022, doi: *10.1109/JBHI.2022.3199206*.

[17] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 11121–11128, Jun. 2023. doi: *10.1609/aaai.v37i9.26317*.

[18] X. Nie, X. Zhou, Z. Li, L. Wang, X. Lin, and T. Tong, "LogTrans: Providing Efficient Local-Global Fusion with Transformer and CNN Parallel Network for Biomedical Image Segmentation," in *Proceedings of the 2022 IEEE 24th Int Conf on High Performance Computing & Communications*, Dec. 2022, pp. 769–776. doi: *10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00128*.

[19] H. Cao, Z. Huang, T. Yao, J. Wang, H. He, and Y. Wang, "InParformer: Evolutionary Decomposition Transformers with Interactive Parallel Attention for Long-Term Time Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 6906–6915, Jun. 2023. doi: *10.1609/aaai.v37i6.25845*.

[20] H. Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence* (AAAI-21), pp. 11106–1115, 2020. doi: 10.48550/arxiv.2012.07436. doi: *10.48550/arXiv.2012.07436*.

[21] Z. Zhang, X. Wang, and Y. Gu, "Sageformer: Series-Aware Graph-Enhanced Transformers for Multivariate Time Series Forecasting," *Proceedings of the International Workshop on Mining and Learning from Time Series*, pp. 1–16, Jul. 2023. doi: *10.48550/arXiv.2307.01616*.

[22] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 22419–22430. url: *https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf*.

[23] S. Liu et al., "Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting," *Proceedings of the International Conference on Learning Representations*, Oct. 2021, pp. 1-20. url: *https://openreview.net/forum?id=0EXmFzUn5I*.

[24] L. Sasal, T. Chakraborty, and A. Hadid, "W-Transformers: A Wavelet-based Transformer Framework for Univariate Time Series Forecasting," *Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications* (ICMLA), Dec. 2022, pp. 671-676. doi: *10.1109/ICMLA55696.2022.00111*.

[25] W. Chen, W. Wang, B. Peng, Q. Wen, T. Zhou, and L. Sun, "Learning to Rotate: Quaternion Transformer for Complicated Periodical Time Series Forecasting," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, in KDD '22. New York, NY, USA: Association for Computing Machinery, Aug. 2022, pp. 146–156. doi: *10.1145/3534678.3539234*.

[26] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting," Jun. 2022. url: *https://api.semanticscholar.org/CorpusID:246430171*.

[27] W. Wang et al., "CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention," *ArXiv*, Oct. 2021. doi: *10.48550/arXiv.2108.00154*.

Transformers for Modeling Long-Term Dependencies in Time Series Data: A Review

BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

NEURAL ENGINEERING DATA CONSORTIUM www.nedcdata.org

S. Thundiyil[1], S.S. Shalamzari[2], J. Picone[2] and S. McKenzie[3]

1. Department of Electronics and Communication Engineering, BMS Institute of Technology and Management, Bengaluru, India
2. The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
3. University of New Mexico Health Sciences, Albuquerque, New Mexico, USA

Abstract

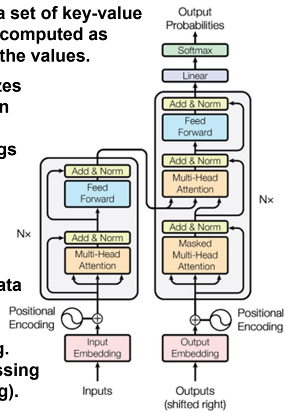
- Long-term dependencies play a crucial role in the analysis of the time series data in applications such as healthcare, climate change, finance etc.
- The emergence of transformer-based technologies such as ChatGPT have demonstrated the potential for analyzing time series data in which modeling of long-term dependencies is crucial.
- In this abstract, we provide a review of various transformer-based models that are used to model long-term context in time series data.
- The combination of self-supervised learning and transformer models is a promising approach to improve the performance and explainability of automatic seizure detection models, especially for continuous EEG data.

Introduction

- Traditional methods such as autoregressive integrated moving average (ARIMA), long short-term memory networks (LSTM), gated recurrent units (GRUs) and recurrent neural networks (RNN) have provided robust frameworks in the analysis of time series data
- These methods have limitations when applied to big data sets and when used to model long-term dependencies.
- Systems based on an attention mechanism leverage positional embedding modules and have been effectively employed in raw EEG data classification related to motor imagery tasks.

A Typical Transformer Architecture

- A transformer architecture relies solely on attention mechanisms that allow the model to process inputs and outputs in parallel.
- The transformer follows an encoder-decoder structure, employing stacked self-attention and point-wise fully connected layers.
- An attention function in the model maps a query and a set of key-value pairs to an output, computed as a weighted sum of the values.
- A transformer utilizes multi-head attention and incorporates positional encodings to maintain information about the order of sequence tokens.
- Transformers can handle raw input data without the need for extensive feature engineering. (minimal preprocessing and postprocessing).



Application Areas

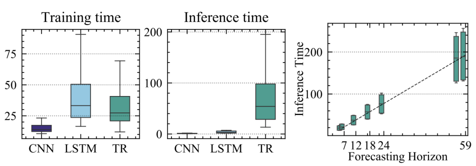
- A summary of various application areas where transformer-based architectures have been successful along with the most popular architectures for each application:

Application of Transformers

Application areas	Popular architectures
Natural Language Processing Machine translation, Text summarization, Question answering, Natural language inference, Sentiment analysis, Named entity recognition, Part-of-speech tagging Text generation, Code generation Code summarization ...	GNMT, Marian, BART, Pegasus, T5, BERT, RoBERT, DistilBERT, GPT, BARD, Codex, CodeGen, CodeT5
Computer vision Image classification, Object detection and segmentation, Image captioning Image retrieval, Image restoration ...	Vision Transformer, Swin Transformer, ConvMixer, DeiT, PVT, Log Trans, Cross former
Audio/speech Automatic speech recognition, Language identification Text-to-speech synthesis, Audio classification, Music generation...	Conformer, Wav2Vec, MuseNet.
Signal processing Time series forecasting, Anomaly detection Classification, Segmentation Denoising, Super-resolution, Signal compression	Informer, Anoformer, WaveTransformer, BENDR, Informer, Sageformer, Autoformer, Pyraformer etc.

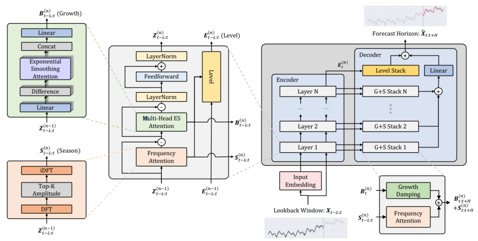
Comparison to Time Series Models

- Transformer models have shown improvements in terms of accuracy, computational efficiency while handling long term dependencies.
- Traditional time series models often have lower computational complexity but may require additional steps for trend and seasonality decomposition, which can increase the overall computation time.
- The self-attention module in standard Transformers has a quadratic time and memory complexity, posing a computational bottleneck for long sequences.
- To address this, models like LogTrans and Pyraformer introduce a sparsity bias in the attention mechanism, while models like Informer and FEDformer utilize the low-rank properties of the self-attention matrix to reduce complexity.
- Traditional models have a limited memory mechanism and can remember and utilize only a fixed number of previous data points. This inherently restricts their ability to capture long-range dependencies effectively.
- While Transformer models offer higher accuracy and better handling of long-range dependencies, they often come with higher computational costs compared to traditional time series models. A comparison of computational efficiency for 12 different time series data sets is shown below:

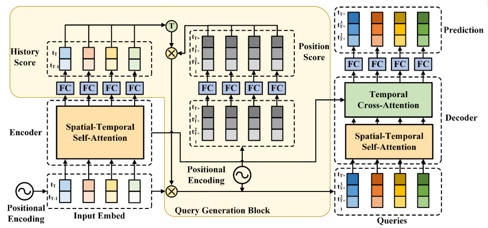


Advancements and Innovations

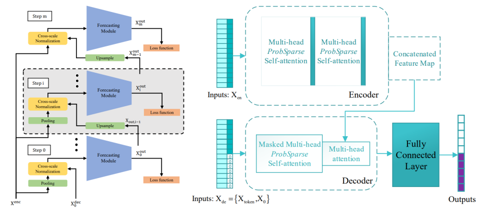
- Standard transformer designs excel in capturing global dependencies, but do not fully exploit the characteristics of time-series data, such as local structures that are better captured by conventional approaches such as convolutional or recurrent architectures.
- Interpretability remains a challenge, raising questions about their trustworthiness and bias.
- Recent innovations in transformer architectures, particularly those focusing on long-term time series forecasting, have introduced several significant advancements.
- Recent developments in attention mechanisms and efficiency enhancements have led to the introduction of more sophisticated time-series forecasting models. ETSFORMER, for instance, leverages exponential smoothing attention and frequency attention to improve efficiency.



- NAST, on the other hand, employs a non-autoregressive architecture with a unique spatial-temporal attention mechanism.



- Innovative Decomposition and Trend Analysis Techniques have been used in TDFormer, Differential Attention Fusion Model and FEDFORMER.
- Enhanced Multiscale and Long-Sequence Forecasting have been implemented in Scaleformer and Informer architectures.



Limitations and Challenges

- **Computational Demands:** Transformers are computationally intensive due to their complex architecture, which can be challenging for long sequence time-series forecasting. They often require high computational resources, particularly for training large models, which can be a limiting factor.
- **Need for Large Datasets:** Transformers typically require large datasets to train effectively due to their numerous trainable parameters. This need for extensive data can be a challenge in scenarios where data is scarce or expensive to acquire, such as many bioengineering or health sciences applications.
- **Overfitting Issues:** There is a risk of overfitting, especially when dealing with time series data that has complex patterns. Overfitting can lead to models that perform well on training data but poorly generalize to new, unseen data.
- **Quadratic Time Complexity:** The self-attention mechanism has a quadratic time complexity with respect to the sequence length, which can be prohibitive for very long time series. This issue limits the scalability of models in certain applications.
- **Handling Long-Range Dependencies:** While transformers are designed to capture long-range dependencies, their effectiveness can vary depending on the nature of the time series data. Application-specific adaptations are required.
- **Context Window Length:** While models learn long-term dependencies, they are limited by temporal coherence and context fragmentation.

Summary

- Compared to traditional models, a transformer architecture can be effective in analyzing long-term dependencies in time series data.
- Long-term dependencies in the data are useful in several applications, such as natural language processing, computer vision, and audio signal processing, as well as in various domains such as healthcare, climate studies, and finance.
- Innovations in transformer models are focusing on efficient ways to model long-term context, which poses a combinatorial problem, and prevents efficient integration of long-term and local constraints.
- This latter point is particularly important in sequential physical signal data such as speech, cardiology or EEG signals.

Acknowledgements

- This material is supported in part by the National Science Foundation under grant no. 2211841. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

Updates to this review will be available on GitHub at https://github.com/sanect/timeSeries_transformers