TECHNISCHE UNIVERSITÄT
ILMENAU

Fakultät für Informatik und Automatisierung

Data-intensive Systems and Visualization Group

Attention is all you need! But which one?

Research Project

Taimur Ibrahim

Betreuer: M. Sc. Dominik Walther

Verantwortlicher Hochschullehrer: Prof. Dr.-Ing. Patrick Mäder

Das Masterarbeit wurde am Datum bei der Fakultät für
Informatik und Automatisierung der Technischen Universität Ilmenau
eingereicht.

Abstract

The Transformer architecture has revolutionized Natural Language Processing (NLP) by leveraging self-attention mechanisms to capture complex relationships between words in a sequence. However, applying Transformers to time-series data presents unique challenges due to their permutation invariance, reliance on fixed positional encodings, and computational complexity for long sequences. This research project aims to investigate the effectiveness of various attention mechanisms in enhancing Transformers' ability to model time-series data. Specifically, we will compare the performance and efficiency of standard attention with at least three alternative mechanisms: Latent Attention, Sparse Attention, Informer's Attention, and Longformer's Attention. By implementing and evaluating these variants on a suitable time-series dataset, we seek to determine which attention mechanism best captures temporal dependencies and improves predictive performance. The project will culminate in a detailed analysis of the models' accuracy, computational efficiency, and potential trade-offs, contributing valuable insights to the application of Transformers in time-series modeling.

Die Transformer-Architektur hat die Verarbeitung natürlicher Sprache (NLP) revolutioniert, indem sie Mechanismen der Selbstaufmerksamkeit nutzt, um komplexe Beziehungen zwischen Wörtern in einer Sequenz zu erfassen. Die Anwendung von Transformers auf Zeitreihendaten stellt jedoch aufgrund ihrer Permutationsinvarianz, ihrer Abhängigkeit von festen Positionskodierungen und ihrer Rechenkomplexität für lange Sequenzen eine besondere Herausforderung dar. Dieses Forschungsprojekt zielt darauf ab, die Wirksamkeit verschiedener Aufmerksamkeitsmechanismen zu untersuchen, um die Fähigkeit von Transformers zur Modellierung von Zeitreihendaten zu verbessern. Konkret werden wir die Leistung und Effizienz der Standardaufmerksamkeit mit mindestens drei alternativen Mechanismen vergleichen: Latent Attention, Sparse Attention, Informer's Attention und Longformer's Attention. Durch die Implementierung und Evaluierung dieser Varianten auf einem geeigneten Zeitreihendatensatz wollen wir herausfinden, welcher Aufmerksamkeitsmechanismus zeitliche Abhängigkeiten am besten erfasst und die

Vorhersageleistung verbessert. Das Projekt gipfelt in einer detaillierten Analyse der Genauigkeit der Modelle, der Berechnungseffizienz und möglicher Kompromisse, die wertvolle Erkenntnisse für die Anwendung von Transformers in der Zeitreihenmodellierung liefern.

(German version generated with DeepL)

Selbstständigkeitserklärung: „Hiermit versichere ich, dass ich diese Masterarbeit selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe. Alle von mir aus anderen Veröffentlichungen übernommenen Passagen sind als solche gekennzeichnet. “

Ilmenau, Datum ____________________________________

Taimur Ibrahim

Contents

Introduction 1

The advent of the Transformer architecture has greatly impacted the field of Natural Language Processing (NLP). It demonstrates exceptional capabilities in capturing complex dependencies within sequences of text through self-attention mechanisms. Introduced by Vaswani et al. in 2017, Transformers have become the backbone of numerous state-of-the-art models in NLP tasks such as machine translation, text summarization, and sentiment analysis [Vaswani et al., 2017]. Its impressive performance in NLP has lead to interest in exploring the application of Transformers in other domains, including time-series modelling.

Time-series data, characterized by sequentially ordered observations over time, is prevalent across various fields including finance, healthcare, and energy management. Accurate modeling and forecasting of time-series data are crucial for decision-making and planning in these domains. Traditional methods like ARIMA and exponential smoothing, as well as machine learning approaches such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), have been extensively used for time-series forecasting. However, these methods often struggle with capturing long-range dependencies and complex temporal patterns inherent in time-series data. Transformers, with their self-attention mechanism, offer a promising alternative due to their ability to model long-range dependencies without being constrained by the limitations of sequential processing found in RNNs and LSTMs. However, the application of Transformers to time-series data poses several challenges. Firstly, the permutation invariance of Transformers necessitates the use of positional encodings to maintain the order of the input sequence, which may not adequately capture temporal relationships crucial for time-series data. Secondly, the global attention mechanism of Transformers, while beneficial for NLP tasks, may not be optimal for time-series data where recent values or local patterns are often more relevant. Lastly, the quadratic complexity of the attention mechanism makes Transformers computationally expensive for long sequences of data.

This project aims to address these challenges by investigating different attention mechanisms within the Transformer architecture to enhance its applicability and performance in time-series modeling. By comparing the standard self-attention mechanism with alternative approaches such as Latent Attention, Sparse Attention, Informer's Attention, and Longformer's Attention, the goal of this project is to determine which mechanism best captures the temporal dependencies in time-series data while balancing predictive performance and computational efficiency.

Background 2

(Are subsections actually recommended?)
(more elaboration needed for the architecture itself? i.e diagrams and the specific equations etc)

2.0.1 Transformer Architecture

The Transformer model is fundamentally built on the concept of self-attention, a mechanism that computes the representation of a sequence by relating each element to every other element in the same sequence. Unlike previous models that relied on recurrent layers to process sequence data sequentially, the Transformer processes all elements simultaneously, significantly improving computational efficiency and reducing training times.

2.0.2 Time Series Analysis and Transformers

Time series analysis involves forecasting future values based on previously observed values, often incorporating complex patterns such as trends, seasonality, and cycles. Traditional methods like ARIMA and Exponential Smoothing have been effective for such tasks but often fail to capture nonlinear relationships as efficiently as machine learning approaches.
Applying Transformers to time series analysis has the potential to leverage their ability to model complex dependencies and interactions across time steps. However, the standard Transformer architecture may not be inherently suited to handle the nuances

of time series data, such as varying importance across time steps and the need for handling long sequences efficiently due to the quadratic complexity of its attention mechanisms.

2.0.3 The Need for Diverse Attention Mechanisms

Given the constraints of the standard attention mechanism in Transformers, exploring alternative attention mechanisms that can reduce computational demands and better capture local and temporal relationships within data is crucial.

Several studies have attempted to adapt the Transformer architecture for time-series analysis. For instance, the Informer model [Zhou et al., 2021b] introduces a Prob-Sparse self-attention mechanism to reduce the quadratic computational complexity of standard attention, making it more feasible for long time-series sequences. Similarly, the Longformer model [Beltagy et al., 2020] employs a combination of local and global attention to handle long documents, which can be analogous to long time-series data. Other approaches include the use of Sparse Attention [Child et al., 2019], which selectively attends to a subset of relevant tokens, thereby reducing computational overhead, and Latent Attention [Dolga et al., 2024], which leverages latent variables to capture underlying patterns in the data. These alternative attention mechanisms offer potential solutions to the limitations of standard Transformers in time-series applications.

Related Work 3

This section reviews the current state of the art in adapting Transformer architectures for time-series modeling.

3.0.1 Transformers in Time-Series Analysis

The pioneering work on Transformers by Vaswani et al. [Vaswani et al., 2017] introduced a model that excels in capturing long-range dependencies through self-attention. However, its direct application to time-series data is non-trivial due to the need for explicit temporal encoding and the high computational cost of processing long sequences. Researchers have proposed several modifications to overcome these limitations.

3.0.2 Informer

The Informer model [Zhou et al., 2021b] addresses the quadratic complexity of the standard attention mechanism by introducing ProbSparse self-attention. This approach selectively focuses on a subset of relevant tokens, significantly reducing computational overhead. Informer also utilizes a distilling operation to handle the redundancy in time-series data, effectively compressing long sequences while retaining essential information. This makes Informer particularly suitable for long-term time-series forecasting.

3.0.3 Longformer

The Longformer model [Beltagy et al., 2020] introduces a combination of local and global attention mechanisms to manage long documents and, by extension, long time-series data. The model employs a sliding window approach for local attention, ensuring that each token attends to its neighbors, while global attention is applied to a selected subset of tokens to capture broader context. This dual attention mechanism balances the need for detailed local information and overall sequence comprehension, making it efficient for long sequences.

3.0.4 Sparse Attention

Sparse Attention mechanisms, such as those proposed by Child et al. [Child et al., 2019], reduce the computational complexity of the standard self-attention by restricting the attention scope to a sparse set of tokens. This approach leverages sparsity patterns to focus computation on the most relevant parts of the sequence, making it scalable to longer time-series. Sparse Attention has shown promise in generating large-scale sequences efficiently, which is crucial for time-series applications involving high-frequency data.

3.0.5 Latent Attention

Latent Attention [Dolga et al., 2024] introduces latent variables to capture the underlying structure of the input data. This method enhances the model's ability to identify hidden patterns and dependencies that are not immediately apparent from the raw data. By incorporating latent variables, Latent Attention can improve the representation of complex temporal relationships, potentially leading to more accurate time-series forecasting.

3.0.6 Other Notable Approaches

Several other models and methods have also contributed to the advancement of Transformers in time-series analysis. For example, the LogSparse Transformer [Li et al., 2020] adapts the attention mechanism to focus logarithmically more on recent tokens than distant ones, aligning better with the typical relevance patterns in time-series data. Additionally, the Reformer [Kitaev et al., 2020] employs locality-sensitive hashing to reduce the complexity of self-attention, making it feasible to process very long sequences efficiently.

3.0.7 Comparison

Each of these models introduces unique solutions to the inherent challenges of applying Transformers to time-series data. Informer and Longformer tackle the issue of sequence length and computational efficiency, while Sparse Attention and Latent Attention offer ways to enhance relevance and capture hidden patterns. However, these approaches also come with trade-offs in terms of implementation complexity and suitability for different types of time-series data.

Methods 4

Experiments

5

Summary 6

Appendix

A

List of Figures

List of Tables

Bibliography

[Beltagy et al., 2020] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[Child et al., 2019] Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

[Dolga et al., 2024] Dolga, R., Cobzarenco, M., and Barber, D. (2024). Latent attention for linear time transformers. *arXiv preprint arXiv:2402.17512*.

[Kitaev et al., 2020] Kitaev, N., Łukasz Kaiser, and Levskaya, A. (2020). Reformer: The efficient transformer.

[Li et al., 2020] Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. (2020). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Wen et al., 2022] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.

[Zhang et al., 2021] Zhang, B., Titov, I., and Sennrich, R. (2021). Sparse attention with linear units. *arXiv preprint arXiv:2104.07012*.

[Zhou et al., 2021a] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021a). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115.

[Zhou et al., 2021b] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021b). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115.