TRANSFORMER EFFECTIVENESS FOR TIME-SERIES FORECASTING

PROJECT PROPOSAL

Marco Lagos & Apoorv Walia

COMP 540: Statistical Machine Learning

Rice University

Fall 2023

Contents

Abstract

This project aims to evaluate the Autoformer, a Transformer-based model, in the context of long-term time series forecasting (LTSF). Our goal is to apply the Autoformer to a wide range of datasets, extending beyond those covered in existing research, to thoroughly assess its effectiveness. In addition, we will conduct comparative analyses against the simple model DLinear, to validate the claims made in recent studies. This investigation is designed to provide a comprehensive understanding of the Autoformer's capabilities in LTSF and to contribute valuable insights into its applicability across various data contexts.

Keywords Transformers · Time Series Forecasting · Autoformer · Linear Models · Autocorrelation

1 Introduction

Long-term time series forecasting (LTSF) has recently witnessed a surge in the application of Transformer-based models, specifically the Autoformer. The Autoformer stands out due to its unique architecture which integrates traditional time series decomposition methods. It separates time series into trend-cycle and seasonality components through a Decomposition Layer, and introduces an auto-correlation mechanism in place of standard self-attention. This design enables the Autoformer to leverage period-based dependencies, enhancing its forecasting capabilities.

Despite its innovation, the effectiveness of Transformer models like Autoformer in LTSF is a subject of debate. A recent study presented in a paper titled "Are Transformers Effective for Time Series Forecasting?" questions the efficacy of Transformers in this domain, citing potential temporal information loss due to their permutation-invariant nature. Conversely, the Autoformer paper published in 2021 "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting" and an independent study Hugging Face argue that Transformers, specifically Autoformer, are indeed effective, providing empirical evidence to counter the former paper's claims.

A key model in this debate is DLinear, a simple linear model that uses the decomposition layer from Autoformer. It decomposes time series data into residual (seasonality) and trend components, each processed through its own linear layer for forecasting. The DLinear model's purported superiority over complex Transformer models in the study is the crux of our investigation.

Our project is centered on exploring these contrasting perspectives by applying the Autoformer and the DLinear models to various datasets and conducting comparative analyses. Through this, we aim to uncover the truth behind these contradicting claims and gain a deeper understanding of the Autoformer's true capabilities in LTSF.

2 Datasets

Our investigation covers a number of mainstream time series forecasting applications: electricity consumption, traffic flow, weather patterns, exchange rate and stock market data in financial markets, to provide a comprehensive evaluation. These datasets, multivariate in nature, are a testing ground to compare the predictive abilities of Autoformer and DLinear models across different contexts. Electricity (5) dataset contains the hourly electricity consumption of 321 customers from 2012 to 2014. Traffic (6) is a collection of hourly data from California Department of Transportation, which describes the road occupancy rates measured by different sensors on San Francisco Bay area freeways. Weather (6) is recorded every 10 minutes for 2020 whole year, which contains 21 meteorological indicators, such as air temperature, humidity, etc. Exchange (6) records the daily exchange rates of eight different countries ranging from 1990 to 2016. Stock market (4) is a collection of daily stock prices of a number of companies from 2005 to 2017.

3 Architecture

3.1 Autoformer

Autoformer, designed for long-term forecasting in applications like extreme weather early warning and long-term energy planning, overcomes the limitations of traditional Transformer-based models in handling complex temporal patterns and long series efficiency. It achieves this through a novel decomposition architecture combined with an Auto-Correlation mechanism, leading to a significant improvement in forecasting accuracy. This section delves into specifics about the Autoformer model; however, for more detailed information on the model see (1).

3.1.1 Time Series Forecasting and Challenges

Time series forecasting is critical in various domains such as energy, traffic, economics, weather, and disease forecasting. Traditional models, including ARIMA and RNNs, have focused on temporal relation modeling. Transformer-based models have shown promise due to their self-attention mechanisms in handling long-term dependencies. However, these models struggle with the complexity and computational demands of long-term forecasting, particularly due to intricate temporal patterns and the quadratic complexity of sequence length.

3.1.2 Decomposition Architecture

Autoformer introduces a progressive decomposition approach to time series analysis. This approach involves separating the series into trend-cyclical and seasonal parts, addressing entangled temporal patterns and highlighting inherent time series properties. This decomposition is vital for capturing trends and peaks in the seasonal part of the series.

Mathematical Formulation

Series Decomposition Block: $X_t = AvgPool(Padding(X))$, where X_t represents the trend component.

Model Input Initialization:

$$X_{ens}, X_{ent} = SeriesDecomp(X_{en}, \frac{I}{2} : I)$$

$$X_{des} = Concat(X_{ens}, X_0),\ X_{det} = Concat(X_{ent}, X_{Mean})$$

3.1.3 Auto-Correlation Mechanism

Autoformer's Auto-Correlation mechanism replaces traditional self-attention with a focus on period-based dependencies and sub-series aggregation. This mechanism is efficient and adept at handling the complexities of long-term time series forecasting.

Mathematical Formulation

Auto-Correlation Mechanism:

$$\tau_1, \ldots, \tau_k = argTop_k(R_{QK}(\tau)),\ \tau \in \{1, \ldots, L\}$$

$$Auto - Correlation(Q, K, V) = \sum_{i=1}^{k} Roll(V, \tau_i) R_{QK}(\tau_i)$$

Encoder and Decoder Equations:

$$X_{en}^{l} = Encoder(X_{en}^{l-1}),\ S_{en}^{l2} = SeriesDecompFeedForward(S_{en}^{l1}) + S_{en}^{l1}$$

$$X_{de}^{l} = Decoder(X_{de}^{l-1})$$

3.1.4 Performance and Evaluation

Autoformer has been extensively evaluated on real-world benchmarks in various domains, consistently outperforming other models in both multivariate and univariate settings. Its robustness and accuracy in long-term forecasting have been demonstrated across different applications, confirming its practical utility.

3.2 DLinear: A Simplified Linear Model for Time Series Forecasting

DLinear, a variant of the LTSF-Linear model, offers a simplified approach to time series forecasting. It is part of a set of linear models designed to challenge the complexity of Transformer-based Long-Term Time Series Forecasting (LTSF) solutions. DLinear, specifically, combines a decomposition scheme with linear layers for efficient forecasting. This section delves into specifics about the DLinear model; however, for more detailed information on the model see (2).

3.2.1 Basic Concept of LTSF-Linear

LTSF-Linear is a basic Direct Multi-Step (DMS) forecasting model that uses a temporal linear layer to directly regress historical time series data for future predictions. This model operates on the principle of weighted sum operations and is represented mathematically as $\hat{X}_i = WX_i$, where $W \in \mathbb{R}^{T \times L}$ is a linear layer along the temporal axis, and $\hat{X}_i$ and X_i are the predicted and input values for each i^{th} variate, respectively. LTSF-Linear is designed to share weights across different variates without modeling spatial correlations.

3.2.2 DLinear Model

DLinear is a specialized variant of the LTSF-Linear model, incorporating a decomposition approach used in Autoformer and FEDformer models, alongside linear layers. It is specifically tailored for time series data that exhibit a clear trend.

3.2.3 Decomposition Scheme

DLinear starts by decomposing the raw data input into two components: a trend component, extracted using a moving average kernel, and a remainder, which is considered the seasonal component. After decomposition, DLinear applies separate one-layer linear models to each of these components. The outputs of these linear layers, corresponding to the trend and seasonal components, are summed to derive the final prediction.

3.2.4 Enhanced Performance

By focusing on the trend component in the data, DLinear enhances the performance of the basic linear model, particularly in datasets with a clear trend. DLinear represents a straightforward yet effective approach for time series forecasting, particularly suitable for data with identifiable trends. It simplifies the forecasting process by leveraging a decomposition strategy combined with linear modeling, offering an efficient alternative to more complex models. The DLinear model demonstrates that even with a basic approach, significant forecasting accuracy can be achieved, especially in time series with prominent trend patterns.

4 Utilizing Pretrained Autoformer Models on Hugging Face

We are employing the pretrained Autoformer models from Hugging Face for datasets related to electricity usage and exchange rates. This approach enables us to rapidly evaluate the model's performance and benchmark it against previously published results. There is a difficulty in utilizing these models given that most of the information involved in their training and publication is not made available.

4.1 Electricity Dataset

Model Setup: Employ the pretrained Autoformer model for electricity data to forecast power consumption trends.

Implementation:

1. Load the pretrained Autoformer model from Hugging Face.
2. Prepare the electricity dataset for testing.
3. Evaluate the model's forecasting accuracy.

4.2 Exchange Rate Dataset

Model Setup: Adapt the model for predicting currency exchange rates.

Implementation:

1. Utilize the pretrained Autoformer model specific to the exchange rate dataset.
2. Conduct forecasting experiments.
3. Analyze the model's precision in predicting future rates.

4.3 Training Larger Models

Parallel to the above auxiliary experiments, we also engaged in training larger models on the data mentioned in the datasets section. This process was progressing slowly, due to availablecomputational power. Our preliminary findings suggest that the statistics reported in the paper for these larger models are accurate. In contrast, replicating the results claimed in the Hugging Face post is proving to be more challenging due to less precise documentation and specifications.

5 Experiments

5.1 Objective

Our primary goal was to address the permutation-invariance in Transformer models' self-attention mechanism to enhance long-term time series forecasting accuracy.

5.2 Design and Methodology

We employed advanced techniques adapted from "Temporal Attention for Language Models" to improve temporal dynamics in Transformer models such as Autoformer.

5.3 Embedding Strategies

Our approach included the implementation of various embedding classes:

- **PositionalEmbedding**: Implements sinusoidal positional embeddings.
- **TokenEmbedding**: Applies a convolutional layer for input token transformation.
- **FixedEmbedding**: Embeds fixed-size categorical inputs.
- **TemporalEmbedding**: Embeds various temporal features.
- **TimeFeatureEmbedding**: Transforms time-related features linearly.
- **DataEmbedding Variants**: Composite embeddings combining value, temporal, and positional embeddings, with specific variants excluding positional and/or temporal components.

5.4 Layer Variation

We experimented with different numbers of encoding and decoding layers to assess their impact on the model's performance.

5.5 Hyperparameter Optimization

A grid search was conducted to optimize the model's hyperparameters, ensuring the best possible configuration for our experiments.

5.6 Learning Rate Adjustment Strategies

We tested various learning rate adjustment strategies to explore their impact on model training:

1. **Type 1**: Rapid halving at each epoch.
2. **Type 2**: Step-wise reduction at predetermined epochs.
3. **Type 3**: Stable rate for the first 9 epochs, then reduced.
4. **Type 4**: Stable rate for the first 14 epochs, then reduced.
5. **Type 5**: Stable rate for the first 24 epochs, then reduced.

6. **Type 6**: Early reduction after 4 epochs.

7. **Type 7**: Multi-phase reduction at epochs 5 and 25.

5.7 Results and Observations

Our experiments revealed that, despite the integration of advanced embedding techniques and temporal attention mechanisms, the simpler DLinear model outperformed our enhanced Transformer models in long-term time series forecasting tasks. The performance of each model across different epochs is illustrated in Figure 1.

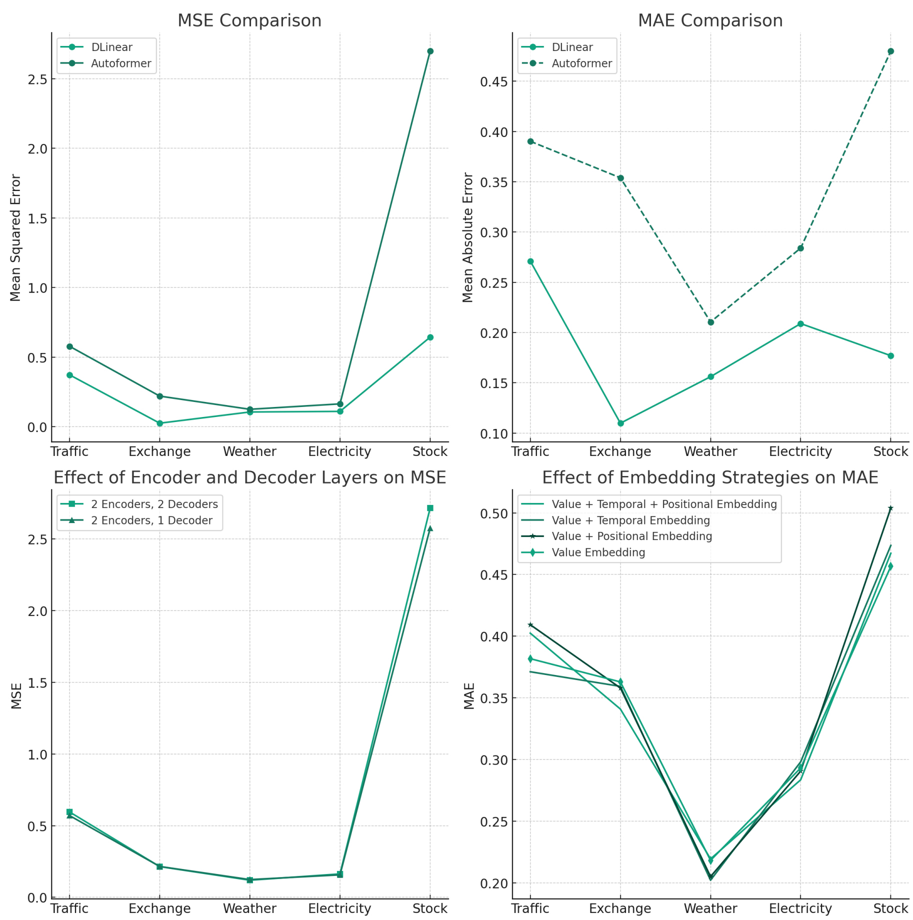

Figure 1: Performance comparison of Transformer models and DLinear model over the datasets.

This graphical representation provides a clear comparison between the various datasets, highlighting the unexpected superiority of the DLinear model in this specific forecasting task.

5.8 Comparison of Prediction Outputs: DLinear Model vs. Autoformer

In our comparative analysis of the DLinear and Autoformer models, we observed distinct characteristics in their prediction outputs. Notably, while the DLinear model exhibited a closer fit to the ground truth in our datasets, the Autoformer seems to demonstrate a superior ability to capture and reflect autocorrelations in the data.

DLinear Model Performance: The DLinear model showed a high degree of accuracy in fitting the available data points. This suggests its effectiveness in capturing the immediate patterns present in the time series, leading to a close alignment with the ground truth in our controlled test scenarios.

Autoformer's Autocorrelation Strengths: In contrast, the Autoformer's predictions, though not as tightly aligned with the ground truth as the DLinear model, displayed a deeper understanding of the underlying data patterns. It demonstrated a more pronounced ability to integrate and leverage autocorrelations, indicating an enhanced comprehension of historical data trends.

Implications for Generalization: This distinction in model behavior leads us to hypothesize that the Autoformer, with its proficiency in handling autocorrelations, may exhibit superior generalization capabilities, especially in scenarios involving longer time series and substantially larger datasets. The Autoformer's ability to account for past data suggests that it could be more adept at predicting future trends when provided with more extensive historical data, a scenario where simpler models like the DLinear might fall short due to their more immediate data fitting approach.

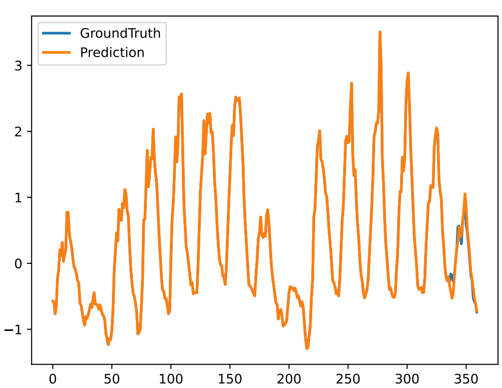

Figure 2: Prediction result of DLinear on the electricity dataset

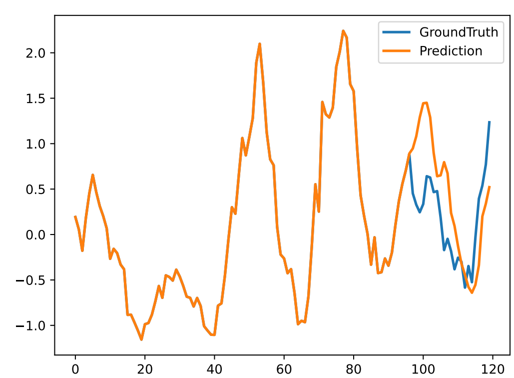

Figure 3: Prediction result of Autoformer on the electricity dataset.

5.9 Limitations and Future Work

Data Size and Transformer Efficiency: One possible limitation of our study is the size of the datasets used. Transformers are known for their data-hungry nature, requiring large amounts of data to effectively learn and generalize. In our experiments, the relatively small datasets might not have been sufficient to leverage the full potential of Transformer models. Future work should focus on experimenting with larger and more diverse datasets, which could provide the depth and complexity needed for Transformers to outperform simpler models.

Autocorrelation and Model Performance: We observed that the auto-correlation features did not significantly enhance the performance of Autoformer models. This could be attributed to the datasets being too simplistic, leading to a scenario where Transformers, due to their complex architectures, might be overfitting. The noticeable disparity between training loss and test loss in our runs supports this hypothesis. Further research should explore the use of more complex and challenging datasets, where the advanced capabilities of Transformers can be more effectively utilized.

Pre-training and Fine-tuning of Autoformers: Another avenue for future research is the pre-training of Autoformers on a large corpus of time series data, followed by fine-tuning on specific tasks. This approach, similar to the methodologies used in natural language processing, could potentially enhance the performance of Autoformers. Training models from scratch on specific tasks might not fully exploit their learning capabilities, whereas pre-training could p on a large corpus of time-series data could provide a more robust and generalized foundation, leading to improved results upon fine-tuning.

Additionally, due to computational and time constraints, certain experiments, including architectural modifications in Autoformer, were not conducted. Future research will focus on exploring these areas for further improvements.

6 Conclusive Thoughts

6.1 Observations and Findings

One of our most important observations was the unexpected superiority of the DLinear model over the more complex Autoformer. This challenges the common perception that complexity in model architecture invariably leads to better performance, especially in time series forecasting. It demonstrates that, in certain scenarios, really simple models can capture the right patterns more effectively.

Our project highlighted the crucial role of time series decomposition in forecasting. Both the Autoformer and DLinear models leverage decomposition, yet their approaches and results are markedly different. This observation underscores the importance of understanding underlying data patterns of the data.

Additionally, the project expanded our understanding of the Transformer architecture (its rise to popularity) and its application in time series forecasting. The integration of auto-correlation in Autoformer was actually intuitive and elegant, providing insights into how Transformer models can be tailored to handle time series data more effectively and any other data it might come across.

Finally, analyzing conflicting studies revealed the nuances in model evaluation. It emphasized clearly the need for comprehensive testing across diverse datasets to validate claims about model performance. And even among top researchers and professionals, there can be a remarkable difference in results depending on details in the approaches.

6.2 Overall Thoughts

This project was a deep dive into the world of machine learning, specifically in the realm of time series forecasting. We gained insights into how different algorithms approach forecasting and the importance of matching model complexity with characteristics in the data.

The most challenging part was reconciling contradictory findings from different studies and understanding why certain models performed better in specific contexts. The most exciting aspect was observing the unexpected performance of simpler models, which was a departure from conventional wisdom.

Conclusively, this project reinforced the notion that in the field of machine learning, there are no one-size-fits-all solutions, and that sometimes, simpler approaches can yield surprisingly effective results.

7 References

- 1. Wu, H., Xu, J., Wang, J., & Long, M. (2022). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. arXiv preprint arXiv:2106.13008.
- 2. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022). Are Transformers Effective for Time Series Forecasting? arXiv preprint arXiv:2205.13504.

- 3. Hugging Face Blog. (2023, June 16). Autoformer. [Online]. Available: https://huggingface.co/blog/autoformer
- 4. Kaggle. (Updated 6 years ago). Stock Time Series Dataset (2005-2017). [Online]. Available: https://www.kaggle.com/datasets/szrlee/stock-time-series-20050101-to-20171231
- 5. Trindade,Artur. (2015). ElectricityLoadDiagrams20112014. UCI Machine Learning Repository. https://doi.org/10.24432/C58C86.6. Guokun Lai. (2017). Multivariate Time Series Data. [Online]. Available: https://github.com/laiguokun/multivariate-time-series-data. [Accessed: 2023].
- 6. Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2022. "Are Transformers Effective for Time Series Forecasting?" arXiv preprint arXiv:2205.13504.
- 7. Guy D. Rosin and Kira Radinsky. 2022. "Temporal Attention for Language Models." arXiv preprint arXiv:2202.02093