DecoFormer:Enhancing CrossFormer’s Performance through DLinear Integration

Zhiyan Li

Abstract

This paper proposes **Decoformer**, an efficient architecture for multivariate time series forecasting (MTS) based on the existing CrossFormer model [3] whose design of Cross-Time Stage (CTS) applies time-invariant Multi-Attention Mechanism. However, considering that MTS forecasting is always time-related, Decoformer removes CTS module and integrates the Decomposition Linear (*Dlinear*) to address this limitation. The resulting hybrid architecture shows stronger forecasting accuracy than Crossformer in long-term prediction with fixed relatively short input length on two real-world benchmark. Moreover, a 1.25x speedup is achieved due to the removal of the CTS module while maintaining performance.

1 Introduction

Multivariate Time Series (MTS) forecasting plays a crucial role in diverse domains, including finance, energy, and healthcare, therefore accurately predicting future trends imposes a tangible impact on decision-making. Recent advances in deep learning, particularly Transformer[1], have demonstrated remarkable potential in capturing dependencies. Among those Transformer-Based models, CrossFormer [3] introduces a novel cross-time attention mechanism to evaluate interactions across different time steps. However, real-world MTS forecasting tasks inherently exhibit time-varying patterns, whereas CrossFormer’s Cross-Time Stage (CTS) employs a time-invariant multi-head attention mechanism.

To fill the gap, we propose Decoformer, an efficient hybrid architecture removing the CTS module and integrating Decomposition Linear (Dlinear) [2] to better capture time-dependent relationships in MTS data. Unlike CrossFormer, which relies on static attention across time steps by applying Two-Stage-Attention (TSA) layer, Decoformer combines Dlinear to harness its strengths in high-accuracy predicting through trend and seasonal components with Cross-Dimension Stage (CDS) which captures dimensional traits in multivariate predicting condition. **The contributions of this paper are:**

1) By replacing CTS with Dlinear, time dependencies are better observed. Particularly, experimental results on two real-world benchmarks show that Decoformer outperforms Crossformer on fixed input length and long-term output length.
2) The removal of the Cross-Time Stage in the Two-Stage Attention Layers results in a 1.25x speedup, making Decoformer more efficient while maintaining accuracy.

2 Method

2.1 DLinear

DLinear Model is a simple yet effective framework for Time Series Forecasting (TSF), designed to address the limitations of complex Transformer-based models, such as high computational overhead and potential overfitting. As shown in Figure 1, this architecture decomposes time series into trend and residual components and employs dual linear layers to model them separately, achieving competitive performance with minimal computational complexity.

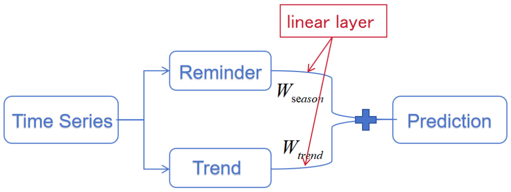

Figure 1: Stages of DLinear Model

Decomposition Layer

A moving average kernel is applied to the input time series $X \in \mathbb{R}^{T \times D}$, where T is the input length of time series, and D is the feature dimension, to separate trend and residual components.

$$X_{trend} = \text{AvgPool}(\text{Padding}(X)), X_{reminder} = X - X_{trend}$$

Dual Linear Layers

$$X^{time}_{T+1:T+\tau} = W_s X_{reminder} + W_t X_{trend}$$

, where W_s, $W_t \in \mathbb{R}^{\tau \times T}$ are learnable projection matrix, and τ,$X^{time}_{T+1:T+\tau}$ are the prediction length and final output.

Computation Complextity

The computation complexity of Dlinear model is $O(DL)$, where L represents for input length of time series. After this stage, seasonal and residual features are caputured in $X^{time}_{T+1:T+\tau}$. We regard this as output for time-related prediction.

2.2 Cross-Dimension Stage

In the CDS, the MTS input is initially transformed into a 2D vector representation using a technique called Dimension Segment-Wise embedding (DSW embedding). This embedding approach allows us to disentangle and segment important features from each dimension while simultaneously encoding the interrelationships between dimensions. By converting the original high-dimensional MTS data into this enriched representation, the model can better analyze cross-dimension dependencies.

Once the feature embedding is complete, we leverage the Multi-Head Self-Attention (MSA) mechanism within the CDS. MSA is particularly well-suited for capturing interactions across different dimensions and time scales, as it allows the model to attend to and focus on specific patterns or features that are important for prediction. Through the application of MSA, we dynamically model the dependencies not only within an individual dimension but also across all dimensions in the input data.

Following the cross-dimension modeling, we construct a Hierarchical Encoder-Decoder (HED) architecture. This architecture is designed to efficiently process the complex relationships embedded in the input data across multiple levels of representation.The basic design of CTS and HED are displayed in the below figure 2. The final predicting output $X^{dim}_{T+1:T+\tau}$ is dimension-related.

The specific implementation details are in the Crossformer[3].

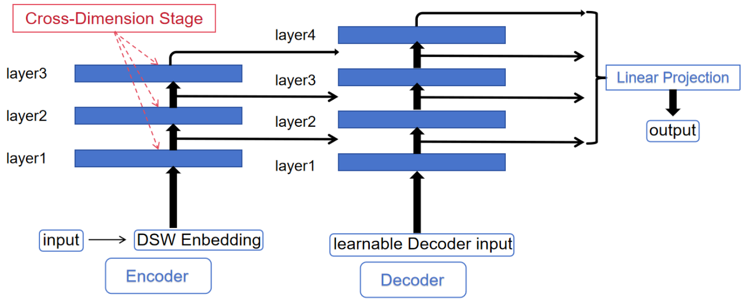

Figure 2: Structure of CTS and HED

2.3 Combination

Since we get two prediction through DLinear and Cross-Dimension Stage, each stands for time-related output and dimension-related output, we introduce a learnable parameter, denoted as α, which serves as a weighting factor to balance the contributions of the two components. The parameter α is trained alongside the rest of the model during the learning process, allowing the system to dynamically adjust the relative importance of time-related and dimension-related outputs based on the specific characteristics of the dataset. Moreover, $\alpha \in [0, 1]$.

$$X_{final} = \alpha X^{dim}_{T+1:T+\tau} + (1-\alpha)X^{time}_{T+1:T+\tau}$$

Decofomer structure are shown in Figure3.

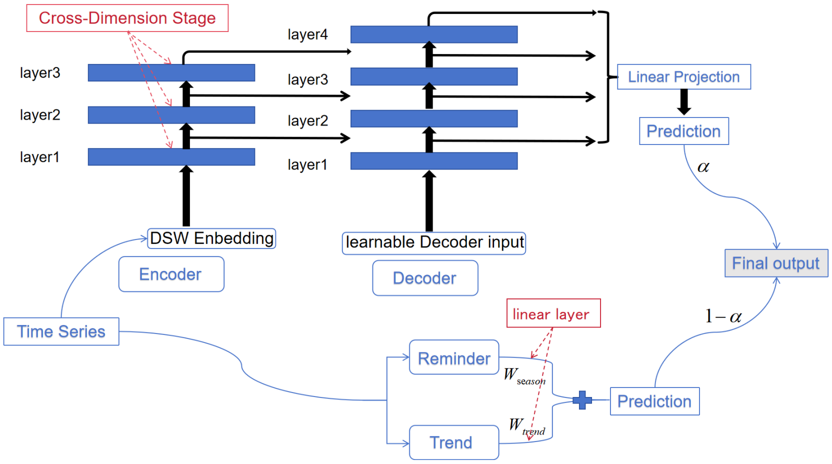

Figure 3: Structure of Decoformer

3 Experiments

3.1 Datasets and Baseline

Decoformer is trained and tested following **1)ETTh1** (Electricity Transformer Temperature-hourly), **2)ETTm1** (Electricity Transformer Temperature-minutely) datasets[4], and we use Crossformer's output of MTS forecasting problems as baseline.

3.2 Main Results

The settings are same as that in Zhou ei al.(2021)[4]: train/val/test has the ratio **0.7:0.2:0.1**. On each dataset, we fix the input length of time series as 96, and change future prediction length to evaluate model's performance.We also set train batch as 32.

		Crossformer		**Decoformer**	
Dataset	Prediction Length τ	MSE	MAE	MSE	MAE
ETTh1	24	0.2972	0.3499	0.3038	0.3653
	48	0.3570	0.4003	0.3452	0.3910
	168	0.4738	0.4737	0.4300	0.4415
	336	0.5866	0.5507	0.4819	0.4739
	720	0.6981	0.6360	0.5112	0.5181
ETTm1	24	0.2290	0.3021	0.2372	0.3065
	48	0.2758	0.3422	0.2767	0.3348
	96	0.3540	0.3915	0.3329	0.3737
	288	0.4741	0.4853	0.4213	0.4411
	672	0.5987	0.5683	0.5478	0.5337

Table 1: MSE/MAE of Crossformer and Decoformer with fixed input $T = 96$ and differrent prediction lengths τ

As shown in Table 1 ,Figure 4 and Figure 5, when the prediction lengths are relatively short ($T = 24, 48$), the MSE and MAE of Crossformer and Decoformer are close in value. However, as the prediction lengths increase, Decoformer achieves significantly lower MSE and MAE compared to Crossformer, highlighting its clear advantage and greater effectiveness in handling long-term forecasting tasks. This demonstrates the robustness and superior design of Decoformer in capturing patterns for extended prediction horizons.

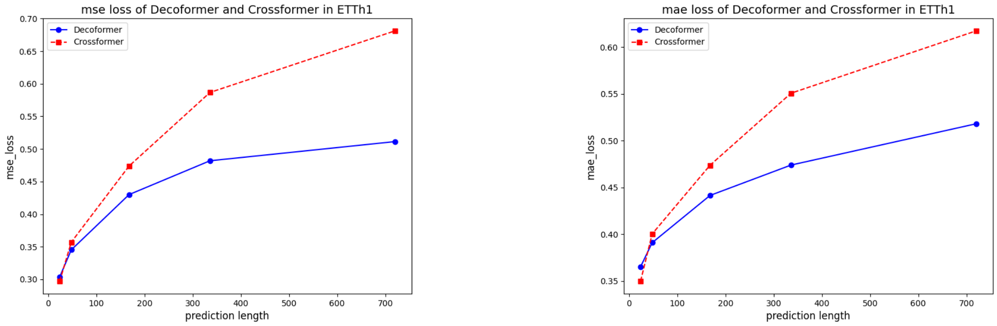

Figure 4: MSE/MAE loss of Decoformer and Crossformer on ETTh1

3.3 Effect of Hyper-Parameters

3.3.1 Segment Length

In the DSW embedding phase, we rearrange the time series into a 2D array according to the segment length. It has been observed that, particularly in the context of long-term forecasting on the ETTh1 dataset, increasing the segment length from 6 to 24 is beneficial for improving performance. The rationale for this improvement could be that a segment length of 24 aligns well with the daily periodicity inherent in the ETTh1 dataset [3].

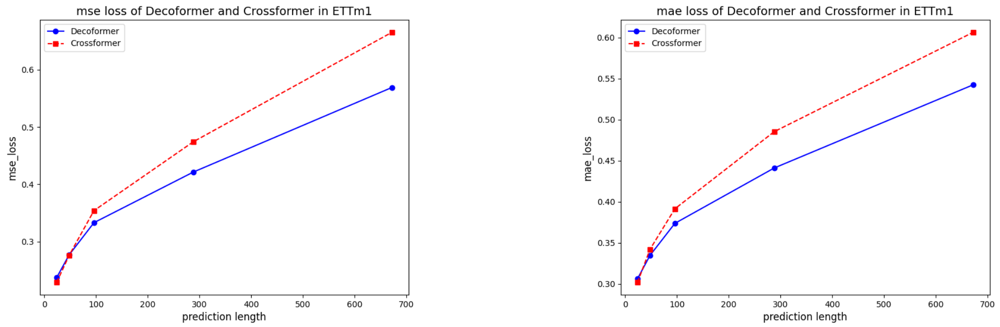

Figure 5: MSE/MAE loss of Decoformer and Crossformer on ETTm1

3.3.2 DLinear Window Size

The DLinear Window Size refers to the length used for pooling operations when processing trend vector X_{trend}. By adjusting this hyper-parameter, we control the granularity of the input data representation fed into the DLinear layer. Larger window sizes aggregate information over broader temporal spans, which can help the model capture long-term dependencies. Conversely, smaller window sizes retain finer-grained details. Determining the optimal window size is essential for balancing computational efficiency and prediction accuracy, especially in long-term forecasting.

We experiment on changing DLinear window size on the same setting (input length $T = 96$, prediction length $\tau = 720$, use ETTh1 as dataset), and calculate the average loss after 5 iterations.The result is shown in Figure 6.

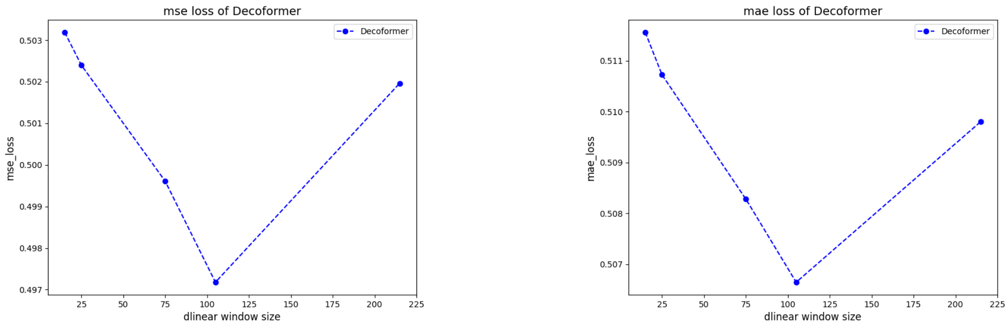

Figure 6: MSE/MAE loss of Decoformerwith different window size

3.3.3 Choice of Hyper-Parameters

After experimenting, hyper-parameters are suggested as follows

1) For short-term prediction($\tau = 24, 48, 72$), set seg_len as 6, $learning\_rate$ as 10^{-4}, e_layers as 3, $dwin\_size$ as 23.
2) For long-term prediction ($\tau = 168, 672, 720$), set seg_len as 24, $learning\_rate$ as 10^{-5}, e_layers as 2, $dwin\_size$ as 105.

3.4 Computational Efficiency Analysis

The total complexity of Cross-Time Stage is $O(DL^2)$, whereas that of DLinear model is $O(DL)$. This reduces the original time complexity of the TSA Layer from $O(DL^2 + DL)$ to $O(DL)$.

While dividing the series into timesteps with length $\frac{T}{L_{seg}}$, Encoder's complexity is $O(D\frac{T}{L_{seg}})$.In long sequence forecasting tasks, we typically increase the length of L_{seg}. Consequently, the computational complexity decreases significantly. This adjustment effectively reduces the computational cost, making the model more efficient for handling long sequences.

Moreover, during both training and inference, we carefully monitored the execution speed of each iteration. Empirical results show that, with the optimized model architecture and algorithm, there is an average speedup of 1.25x per iteration compared to the original implementation. This significant improvement in efficiency highlights the advantages of our optimization strategy, enabling faster training and inference while maintaining strong performance, particularly for long sequence forecasting tasks.

4 Conclusion and Future work

In this paper, we introduced Decoformer, an efficient and accurate hybrid architecture tailored for multivariate time-series (MTS) forecasting tasks. By addressing the limitations of time-invariant mechanisms in existing models like CrossFormer, Decoformer integrates Decomposition Linear (Dlinear) to capture dynamic time-dependent relationships and better model trend-seasonal interactions. Additionally, the removal of the Cross-Time Stage (CTS) simplifies the architecture, resulting in significant computational efficiency with a 1.25x speedup while preserving high forecasting accuracy. Experimental results on real-world benchmarks demonstrate Decoformers superior performance in both short-term and long-term forecasting scenarios, outperforming CrossFormer in predictive accuracy. By effectively combining the strengths of Dlinear for decomposed feature learning with the Cross-Dimension Stage (CDS) for dimensional interaction modeling, Decoformer provides a robust solution for complex MTS forecasting applications across domains such as finance. We hope this approach accelerates progress in time-series forecasting and inspires further advancements in time-varying pattern analysis.

We also analyzes the limitation of this work.**1) Limited Improvement in Short-Term Forecasting**.One notable limitation of Decoformer is its relatively small improvement in short-term forecasting compared to CrossFormer. This might be attributed to the model's design focus on decomposable components such as trends and seasonality, which are more prominent in long-term patterns but less pronounced in short-term intervals. **2) High dependency on the characteristics of the dataset**.Dlinear makes predictions by capturing trends and patterns in the dataset. However, if the trends within the data are not significant or apparent, the predictive performance of the model may weaken significantly. Additionally, if the data exhibits abnormal fluctuations or unpredictable noise over a certain period, DLinear's performance can also be adversely affected, making it difficult to generate accurate predictions. This indicates that DLinear is better suited for datasets with clear trends and relatively stable patterns. Therefore, in practical applications, implementing appropriate data preprocessing methods to reduce noise or abnormal changes within the dataset will be a crucial approach to improving the prediction effectiveness of the DLinear model.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[2] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022.

[3] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *CoRR*, abs/2012.07436, 2020.