Exploring Spectral Bias in Time Series Long Sequence Forecasting

Kofi Nketia Ackaah-Gyasi*
kofinketia.ackaahgyasi01@utrgv.edu
The University of Texas Rio Grande Valley
Edinburg, TX, USA

Sergio Valdez*
sergio.valdez02@utrgv.edu
The University of Texas Rio Grande Valley
Edinburg, TX, USA

Yifeng Gao
yifeng.gao@utrgv.edu
The University of Texas Rio Grande Valley
Edinburg, TX, USA

Li Zhang
lzhang18@gmu.edu
George Mason University
Fairfax, VA, USA

ABSTRACT

Transformers have achieved great success in the task of time series long sequence forecasting (TLSF) in recent years. However, existing research has pointed out that over-parameterized deep learning models are in favor of low frequency and could be difficult to capture high-frequency information for regression fitting task, named spectral bias. Yet the effect of such bias on TLSF problem, an auto-regressive problem with a long forecasting length, has not been explored. In this work, we take the first step to investigate the spectral bias issues in TLSF task for state-of-the-art models. Specifically, we carefully examine three different existing time series Transformers on the task of TLSF with both synthetic and real-world data and visualize their behavior on spectrum. We show that spectral bias exists in the problem of TLSF. Surprisingly, our experiment demonstrated that the model bias behavior, whether it favors at high or low frequencies, is heavily influenced by the model design of the individual Transformer.

KEYWORDS

time series forecasting, spectral bias

ACM Reference Format:
Kofi Nketia Ackaah-Gyasi, Sergio Valdez, Yifeng Gao, and Li Zhang. 2023. Exploring Spectral Bias in Time Series Long Sequence Forecasting. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (KDD' 23).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Time series long sequence forecasting (TLSF) is one of the most important problems in the field of time series data mining and it has wide applications in different domain such as stock price forecasting [7], traffic flow[1, 4], and electricity consumption [9, 16]. In those applications, it is crucial to have precise and high-resolution automated time series forecasting. For example,

- In the financial market, accurate stock forecasting can be very useful for companies and investors, leading to better long-term financial decisions;
- Forecasting the flow of city crowds would help the government plan ahead for city infrastructure and manage public safety;
- Accurate electricity consumption forecasting provides necessary guidance on planning and allocating energy and resources to small households ahead of the time.

Meanwhile, deep neural network achieves remarkable success in time series forecasting. Compared to traditional algorithms, deep learning-based methods, such as Long Short-Term Memory unit (LSTM) [5], outperforms traditional method such as ARIMA due to the ability to capture nonlinear behavior [10]. However, these methods have limited capability on predicting long sequence [15]. Recently, various Transformer-based methods were introduced to TLSF taks due to the success of the original Transformer in capturing long-term dependencies in text translation and significant advance the state-of-the-art TLSF performance.

Despite the great performance, recent research pointed out spectral bias issues in existing over-parameter deep learning models. Such bias could cause the models to strongly favor certain type of characteristics in the data, while causing unstable performance or even failure in capturing the true relation between the input and output of the data [12]. In this work, we would like to systematically investigate and evaluate the spectral bias on Transformer TLSF models. While the original spectral bias is only on classical regression problem, we focus on auto-regression with a long sequence output. In addition, our evaluation focus on time series Transformers while previous work use CNN model. In summary, our key contributions are as follows:

- We take the first step to invest spectral bias in a new problem, which is time series long sequence forecasting problem, with a auto-regressive setting.
- Through carefully examining three different type of existing Transformer models on the task of time series long-sequence forecasting, We show that the spectral bias exists in TLSF problem in both synthetic data and real-world data.
- We observe empirically that the bias behavior occurring at high or low frequencies is heavily influenced by the model design of the individual Transformer.

*Both authors contributed equally to this research.

KDD' 23, Aug 06–10, 2023, Long Beach, CA

ACM ISBN 978-1-4503-XXXX-X/23/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

Time Series Transformers. Recently, Transformer [11] has been proposed with the computation relying entirely on performing pairwise attention on data sequences. It has good parallel capability, and can better model long-term dependencies existed in the sequences. Li et al. [6] proposed Log-Sparse Transformer to alleviate the memory cost problem caused by full attention between layers. Zhou et al. [14] proposed Informer by learning the location of attention based on a max-mean criterion. Wu et al. [13] proposed Autoformer by integrating a decomposition with an auto-correlation mechanism to replace the self-attention module in Transformer. Zhou et al. [15] proposed FEDformer by introducing Fourier and Wavelet-based attention module to integrate frequency information. Although these methods improve the performance of TLSF problem, there is inadequate comprehensive study on the behavior of current models in spectrum performance.

Bias in Deep Learning Models. While the over-parameterized deep learning models have achieved remarkable success in different domains, recent research shows that these models tend to capture superficial features and bias towards some easy-to-learn features. Most existing studies focus on image domain tasks such as image classification and object recognition. Geirhos et al. [3] pointed out convolutional neural networks (CNN) bias towards capturing texture rather than shape to perform image classification. Wang et al. [12] discovered color bias in deep learning models and used gray-level information to improve the robustness. Choi et al. [2] found context bias in activity classification in image data. All of the above work rely on some image specific features and they are not designed for evaluating bias in time series. The most closely related work to our task is spectral bias by Rahaman et al. [8]. They used Fourier analysis to analyze the result of deep learning models for regression, and found that these models are biased towards low frequency functions. However, their problem is different from our problem in this paper. Firstly, TLSF is a sequence-to-sequence (Seq2seq) auto-regression problem, whereas they only consider fitting a non-linear function to the input. Moreover, they are working on multilayer perceptron (MLP) and we are working on Transformer-based forecasting models, which has more parameters and allow more non-linear fitting.

To the best of our knowledge, there is no study under the auto-regression setting in time series data. It is unclear whether the spectral behavior for transformer-based forecasting models would have the same property as regular function fitting.

3 PRELIMINARIES

In this section, we describe fundamental concepts related to time series and our problem statement.

3.1 Definitions

Univariate Time Series $X = [x_1, ..., x_n]$ is a set of real-valued observations ordered by discrete time step j where $j = 1, 2, \cdots, n$.
Multivariate Time Series $X = X^1, X^2, ..., X^D$ is a set of D co-evolving single dimensional real-valued time series X_i.
Sliding Window $X_{t,L}$ of a multivariate time series X is a contiguous set of points in each dimension of time series X starting from position t with length L, where $1 \leq p \leq T - L + 1$. Typically $L \leq n$.

Table 1: Table 1: Notation for this paper

Notation	Discription
X_t^i	i^{th} dimension of time series vector at time step t
n	Length of time series X
D	Number of dimension of time series X
L_{in}	input length of time series X
L_{out}	prediction length of time series X
$X_{t,L}$	Sliding starting at t of length L

3.2 Problem Statement

Time Series Long Sequence Forecasting Given a D dimensional time series X at time step T, at every time stamp, the task of **time series forecasting** will take an input of historical fixed sliding window of time series segment of length L_{in}, and predict a fixed future sliding window of time series segment of L_{out}. Specifically, we would like to learn a mapping function f from $\mathcal{R}^{L_{in}} \rightarrow \mathcal{R}^{L_{out}}$, such that $f(X_{t-L_{in}+1,L_{in}}) = X_{t+1,L_{out}}$. In the problem of TLSF, L_{out} is typically long up to a few hundred points.

4 PROPOSED EVALUATIONS

4.1 Spectral bias

Next we describe our bias evaluation strategy. Given a forecasting outcome $\bar{y}_i$ produced through model $f(.)$ in each epoch, instead of simply evaluating the performance through difference in time domain, we evaluate it in the frequency domain. Specifically, following the experiment setting in [8], we first compute energy under frequency-domain through Fast-Fourier-Transform i.e.:

$$E(f) = |FFT(\hat{y})|^2 \tag{1}$$

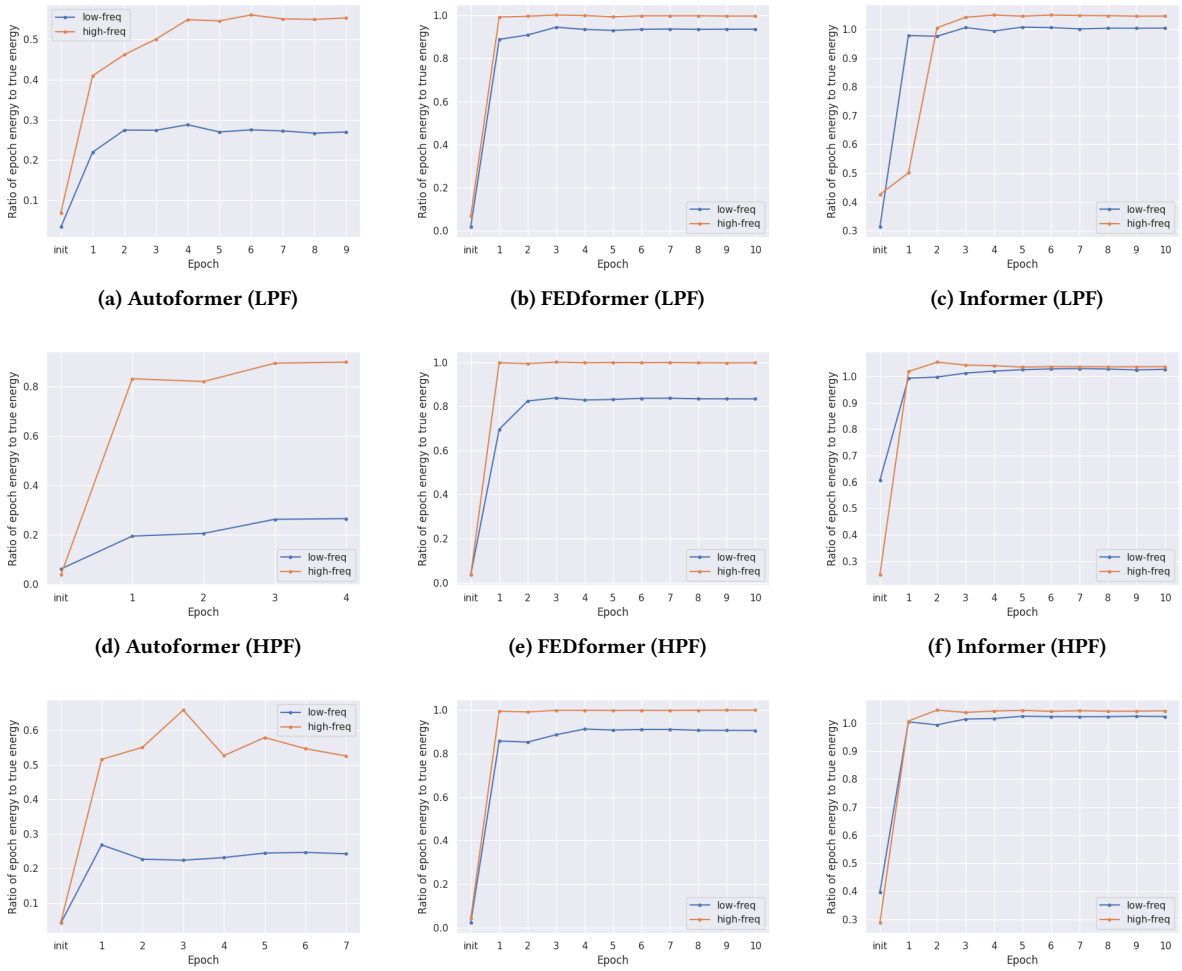
where $\hat{y}$ is the forecasting sequence, f indicates the corresponding frequency, and $E(.)$ indicates the corresponding energy.

In this paper, we will evaluate the spectrum difference in two ways. In the synthetic dataset that we know the ground-truth frequency, f_{gt}, our evaluation will be the ratio between actual energy and the predicted energy:

$$ratio = E'(f_{gt})/E(f_{gt}) \tag{2}$$

where $E'(f_{gt})$ is the energy in frequency f_{gt} produced by the forecasting sequence $\hat{y}$ and $E(f_{gt})$ is the actual energy in the ground truth. Intuitively, to evaluate the performance of forecasting on spectrum, we wanted to take a look at how the energy of an iteration's time series forecasting at a specific frequency compares with the true energy corresponding to the same frequency. This means that if the ratio is equal to 1, then the prediction is accurate, if the ratio is greater than one, the prediction is an overestimate, and if the ratio is less than one, then the prediction is an underestimate.

For the purpose of studying and fully controlling the dataset, we use both quantitive analysis and real-world case study to demonstrate our findings. We first run different Transformer models to get the output, then we visualize the difference between predicted and actual spectrum in different experiment settings.



(a) Autoformer (LPF) (b) FEDformer (LPF) (c) Informer (LPF)

(d) Autoformer (HPF) (e) FEDformer (HPF) (f) Informer (HPF)

(g) Autoformer (EPF) (h) FEDformer (EPF) (i) Informer (EPF)

Figure 1: Ratio of prediction's energy to true energy across the respective epochs for Autoformer, FEDformer and informer transformers in synthetic data

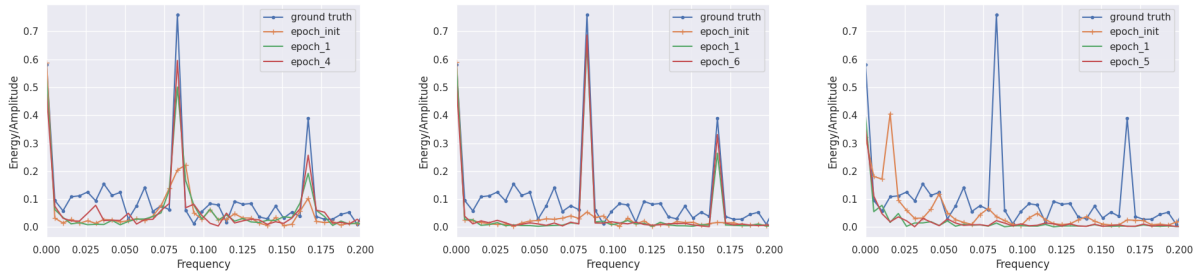
4.2 Time Series Transformers

We evaluate three recent state-of-the-art time Transformer models to evaluate their spectral bias:

- **Informer** [14]. Informer uses a sparse self-attention module which only updates attention weights on top $U = log(L_{in})$ max-mean positions.
- **Autoformer** [13]. Autoformer replaces the self-attention unit with an auto-correlation mechanism and integrates a seasonal decomposition component after each auto-correlation unit. Intuitively, Autoformer considers aligning the peak of a query and keys together.
- **FEDformer** [15]. FEDformer performs attention on frequency domain through Fourier transformation and Wavelet transformation. The goal of the frequency attention is to enhance the learning in high frequency and improve the overall performance of the model.

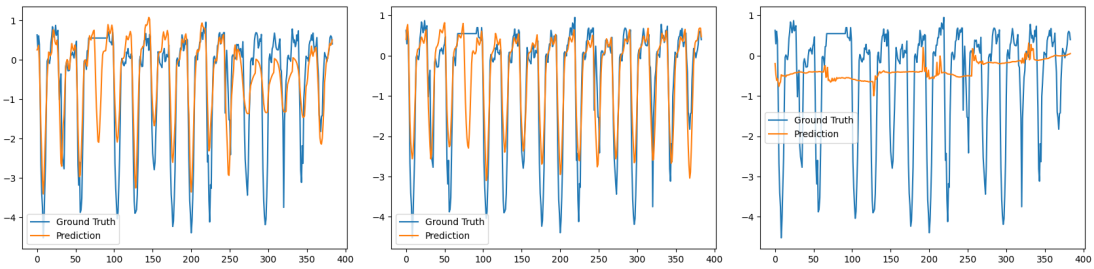
5 EXPERIMENTS AND RESULTS

In this section, we evaluate spectrum bias of Autoformer, FEDformer, and Informer models on each of the three cases of synthetic data. We follow the implementation of FEDformer [15] and use their default setting for all benchmarks to perform our experiments. Across all the models, we use two layers of encoder and one layer of decoder, and set the model embedding size as 512. We use a consistent length across the synthetic and real-world data, the input sequence length L_{in} is set to 256 and the prediction length L_{out} is 384. For all our experiments we use Google Colab cloud GPU T4 of 16GB.



(a) Autoformer (b) FEDformer (c) Informer

Figure 2: Spectrum visualization per initial, first and last epoch for for Autoformer, fedformer, and informer transformers in real world data



(a) Autoformer (b) FEDFormer (c) Informer

Figure 3: Visualization of last epoch in the Time domain

5.1 Dataset

We evaluate the three Transformer models in both synthetic and real-world time series. The synthetic time series is generated by:

$$T(x) = A_1 \sin(\frac{1}{2}x + \pi) + A_2 \sin(2x - \pi) \tag{3}$$

where we construct the following datasets to cover three different cases:

- Low-Frequency Priority (LPF): The energy in the low frequency is two times that of the high frequency (i.e. $A_1 = 2A_2$).
- High-Frequency Priority (HPF): The energy in the high frequency is two times that of the low frequency (i.e. $A_2 = 2A_1$).
- Equal Frequency Priority (EPF): The energy in both frequencies is equal (i.e. $A_1 = A_2$).

In addition, we also evaluate spectral bias on ETTh data, which is a real-world data from an electric power plant in China and is a common benchmark data in the task of TSLF [14].

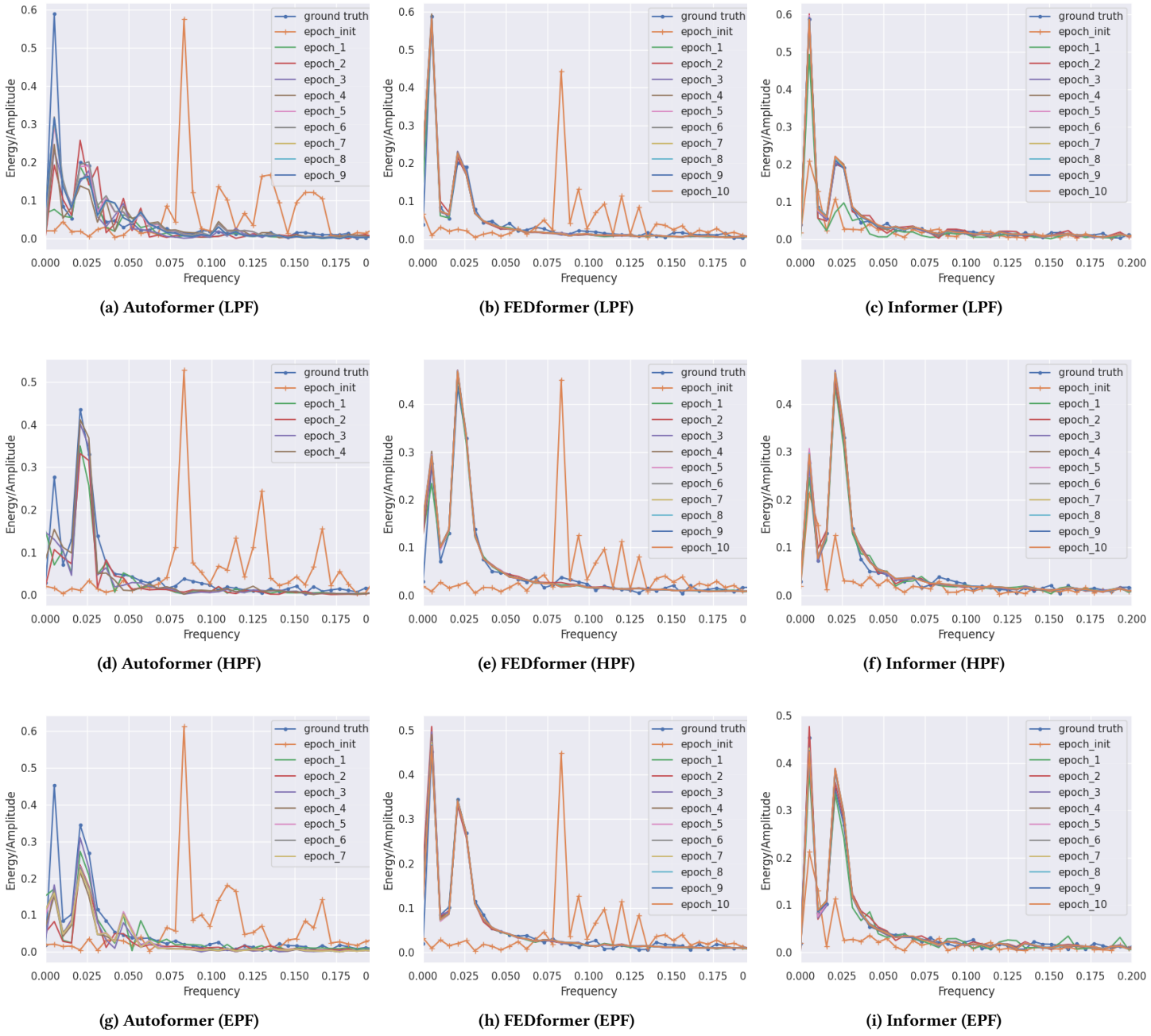
5.2 Predicting Energy Performance in Ground Truth Frequency

We first test the average spectral prediction performance measured on Eq. 2 with the ground truth high frequency and low frequency on our three synthetic data. Figure 1 shows the various visualizations of trend of energy ratios from initial epoch to final epoch for different Transformer models in different synthetic data. The orange line represents the predicted ratio of the high ground truth frequency and the blue line represents the predicted ratio of the low ground truth frequency.

Figure 1(a-c) illustrate energy ratio graphs of Autoformer, FEDformer and Informer models respectively on low frequency priority synthetic data. Figure 1(d-f) represent the same evaluation on high frequency priority data, and figures 1(g-f) represent the result on the equal frequency priority data. According to the figures, Autoformer fits high-frequency information quite good but fits low frequency information with much difficultly. FEDformer has similar observation, but it fits low frequency much better than Autoformer. Different from Autoformer and FEDformer, Informer, which does not use auto-regressive characteristic or frequency information, fits low frequency information easier than high frequency information. Overall, we found the frequency driven Transformers (Autoformer, FEDformer) is biased towards high-frequency and Informer is biased towards low frequency.

5.3 Visualize the Instance Behavior on Synthetic Data

In this experiment, we visualize the behavior of different Transformers in capturing different spectrum at instance level. We pick an arbitrary instance of 100 in testing data and plot the spectral graph for each epoch prediction result until it converges for all three Transformers, and compare with the ground truth spectrum. Figure 2 shows our result of spectrum visualization per epoch for



(a) Autoformer (LPF) (b) FEDformer (LPF) (c) Informer (LPF)

(d) Autoformer (HPF) (e) FEDformer (HPF) (f) Informer (HPF)

(g) Autoformer (EPF) (h) FEDformer (EPF) (i) Informer (EPF)

Figure 4: Spectrum visualization per epoch for Autoformer, FEDformer, and informer transformers in (a-c): LPF synthetic data, (d-f):HPF synthetic data and (g-i) EPF synthetic data

Autoformer, FEDformer and Informer in low-frequency priority, high-frequency priority and equal-frequency priority cases. We can see that Autoformer captures low frequency in a much slower and less stable fashion compared to FEDformer and Informer. Informer captures high-frequency information better and faster when high frequency has higher energy but can be slower in low-frequency priority case. FEDformer is the best in capturing different frequencies in this case regardless of the frequency energy. This is not surprising as the design of FEDformer uses frequency-based attention and theoretically, should work better in our synthetic data.

5.4 Case Study in Real-world Data

Figure 2 and Figure 3 show the prediction of an instance by different Transformer models on time domain and the converted spectrum graph for ETTh data [14]. We arbitrarily picked an instance at the first dimension and observed a different result from the synthetic data. In this instance, Autoformer works well in the first half of the instance in low frequency, but does not work well in the second half of the instance in both high frequency and low frequency. We can see FEDformer performs well in all three major peaks in frequency. Compared with Autoformer, FEDformer has a better power in fitting high frequency. Meanwhile, we see a bad failure in Informer, which is still able to capture a good amount of low frequency, but could barely capture any information in high frequency as shown in both time and frequency domain figures. In a nutshell, FEDformer performs the most consistent among all three Transformers, while Informer is bias towards low frequency and has worst ability to capture high frequency among all three.

6 CONCLUSION AND FUTURE WORK

In this work, we systematically investigate and evaluate the spectral bias on three different Transformer TLSF models. We examine the models in carefully designed synthetic data to compare the spectral bias, and we provide real-world case study and visualization on the spectrum. The bias behavior of the model, whether it favors at high or low frequencies, is heavily influenced by the model design of the individual Transformer. Our work is a first step in investigating spectrum bias issues in TLSF task to improve the usability of deep forecasting models. In the future, we would like to propose new solutions through new models to mitigate this issue.

REFERENCES

[1] Ling Cai, Krzysztof Janowicz, Gengchen Mai, Bo Yan, and Rui Zhu. 2020. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. *Transactions in GIS* 24, 3 (2020), 736–755.

[2] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[3] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

[4] Amir Ghaderi, Borhan M Sanandaji, and Faezeh Ghaderi. 2017. Deep forecast: deep learning-based spatio-temporal forecasting. In *ICML Time Series Workshop*.

[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[6] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*. 5244–5254.

[7] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971* (2017).

[8] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International Conference on Machine Learning*. PMLR, 5301–5310.

[9] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* (2019).

[10] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2018. A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 1394–1401.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[12] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. 2019. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256* (2019).

[13] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[14] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[15] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*. PMLR, 27268–27286.

[16] Thierry Zufferey, Andreas Ulbig, Stephan Koch, and Gabriela Hug. 2017. Forecasting of smart meter time series based on neural networks. In *Data Analytics for Renewable Energy Integration: 4th ECML PKDD Workshop, DARE 2016, Riva del Garda, Italy, September 23, 2016, Revised Selected Papers 4*. Springer, 10–21.