An Adaptive Temporal Attention Mechanism to Address Distribution Shifts

Sepideh Koohfar
Department of Computer Science
University of New Hampshire, Durham, NH
Sepideh.Koohfar@unh.edu

Laura Dietz
Department of Computer Science
University of New Hampshire, Durham, NH
Dietz@cs.unh.edu

Abstract

With the goal of studying robust sequence modeling via time series, we propose a robust multi-horizon forecasting approach that adaptively reacts to distribution shifts on relevant time scales. It is common in many forecasting domains to observe slow or fast forecasting signals at different times. For example wind and river forecasts are slow changing during drought, but fast during storms. Our approach is based on the transformer architecture, that across many domains, has demonstrated significant improvements over other architectures. Several works benefit from integrating a temporal context to enhance the attention mechanism's understanding of the underlying temporal behavior. In this work, we propose an adaptive temporal attention mechanism that is capable to dynamically adapt the temporal observation window as needed. Our experiments on several real-world datasets demonstrate significant performance improvements over existing state-of-the-art methodologies. The code for reproducing the results is open sourced and available online[1].

1 Introduction

Time series forecasting is a vital problem across many domains such as economics [8], retail [4], and healthcare [10] just to name a few which benefits from robust sequence modelling approaches. The temporal behavior of a time series often changes over time, sometimes slowly and other times rapidly which results in distribution shifts. Building a robust and adaptive model in response to such dynamic shifts can be challenging. Transformers [15] have gained popularity in modeling the temporal behavior of a time series. However, the basic attention mechanism in transformers estimates the similarity based on a point-wise vector of the query and key, each representing individual time steps, thereby ignoring the underlying distribution of the surrounding temporal context, we find that adding multiple layers does not fix this insufficiency, see Section 4.1. Lie et al [9] suggest to integrate temporal context by using a convolutional neural network (CNN) to inform the attention mechanism with temporal information. However, our experimental results demonstrate a weakness in this approach, as it is based on a single fixed-length temporal window, which limits the degree of flexibility to respond to distribution shifts on different time scales. We dive into this problem and investigate the importance of models that can adaptively adjust the attention's window size and how this leads to a better forecasting model.

2 Problem Definition

Given the input data prior to time step t_0, the task is to predict the variables of interest for multiple steps into the future from t_0 to $t_0 + \tau$. Given the historical observations denoted

[1] https://github.com/SepKfr/Eff_pattern_matching.git

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

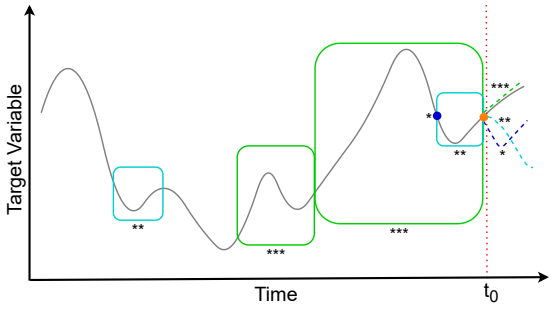

Figure 1: A synthetic example of predicting the future after time step t_0 given the preceding data. The attention mechanism identifies the most similar points for the query at time step t_0, (depicted in orange), and predicts a similar trajectory observed in keys. In the case of point-wise attention this would be the dark blue point (denoted as *) which would result in an erroneous forecast following the dashed dark blue line (denoted as *). In the case of temporal attention with a fixed-length temporal context, the light blue rectangle (denoted as **) is determined to exhibit the most similar behavior which also results in an erroneous forecast following the dashed light blue line (denoted as **). In the case of our adaptive temporal attention the similarity is apparent on different time scales, the big and small green rectangles (denoted as ***), which would result in an accurate forecast depicted in dashed green line (denoted as ***).

as $\boldsymbol{x}_{0:t_0} = [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t_0}]$, we predict the variables of interest for the next τ steps denoted as $\boldsymbol{y}_{t_0:t_0+\tau} = [\boldsymbol{y}_{t_0}, \ldots, \boldsymbol{y}_{t_0+\tau}]$. Where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$, and $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$.

3 Related Work

One of the most dominant classical time series forecasting methods is autoregressive integrated moving average (ARIMA) [3]. ARIMA and other traditional models assumes that the time series is stationary, therefore are not suitable for modeling distribution shifts. Recurrent neural networks (RNNs) have been widely used to model the temporal behavior of time series [12, 2, 11]. The DeepAr [13] model uses long short-term memory networks (LSTMs) to generates parameters of a Gaussian distribution. Transformers have shown superior performance in modeling temporal behavior compared to RNNs [9, 5]. To enhance the forecasting quality several works benefit from employing a CNN layer to integrate the temporal context [8, 9, 14]. Recently, new approaches developed more generalized alternatives to the attention mechanism in transformers to obtain more robust models, from which Autoformer [6] and Informer [16] have demonstrated the best performance. We compare our proposed model to these methods in the experimental section.

4 Methodology

4.1 Background: Point-wise Attention

Given time series data, a single-layer transformer with point-wise scaled dot-product attention predicts the output $\boldsymbol{y}_i$ at time step i as $\boldsymbol{y}_i = \texttt{softmax}(\sum_{j \leq i} \boldsymbol{a}_{ij} \boldsymbol{x}_j)$ with attention $\boldsymbol{a}_{ij} = (\boldsymbol{q}_i^\intercal \cdot \boldsymbol{k}_j / \sqrt{d})$. Typically the query and key vectors are derived via projection $\boldsymbol{q}_i = \texttt{proj}(\boldsymbol{x}_i)$, and $\boldsymbol{k}_j = \texttt{proj}(\boldsymbol{x}_j)$ using two different multi-layer perceptron-style projections of inputs to vectors with dimensionality d. We call this attention point-wise as it estimates the similarity between query $\boldsymbol{q}_i$ and key $\boldsymbol{k}_j$ based on information at time step i and j, and this does not inform the attention mechanism with the underlying temporal context that the query $\boldsymbol{q}_i$ and key $\boldsymbol{k}_j$ are representing. This point-wise attention overlooks the temporal behavior required for more robust predictions, see Figure 1. The canonical approach to integrate the temporal context is to use a multi-layer model. The hypothesis is that from a previous layer, the temporal context can be absorbed into the queries and keys. However, none of the layers have an innate understanding of temporal context, which is a hurdle to overcome during training.

Table 1: Results summary in MSE and MAE of all methods on three datasets. Lower MSE and MAE indicate a more accurate model. Where ▼ denotes significant deterioration compared to our model using a paired-t-test at $p \leq 0.05$.

Dataset	Horizon	Metric	**Ours**	Autoformer	Informer	CNN-trans	Transformer	LSTM	ARIMA
Traffic	24	MSE	**0.404**	0.422▼	0.513▼	0.610▼	0.903▼	0.524▼	2.109▼
		MAE	**0.289**	0.304▼	0.513▼	0.492▼	0.723▼	0.355▼	1.198▼
	48	MSE	**0.433**	0.438▼	0.501▼	0.835▼	1.233▼	0.634▼	2.600▼
		MAE	**0.333**	0.346▼	0.391▼	0.616▼	0.923▼	0.461▼	1.276▼
	72	MSE	**0.408**	0.426▼	0.496▼	1.191▼	0.966▼	0.625▼	2.657▼
		MAE	**0.333**	0.356▼	0.496▼	0.915▼	0.770▼	0.491▼	1.278▼
	96	MSE	**0.404**	0.444▼	0.431▼	0.824▼	1.039▼	0.635▼	2.413▼
		MAE	**0.338**	0.362▼	0.352▼	0.634▼	0.809▼	0.485▼	1.255▼
Air Quality	24	MSE	**0.807**	0.838▼	0.830▼	0.999▼	1.010▼	1.002▼	1.265▼
		MAE	**0.774**	0.785▼	0.788▼	0.874▼	0.878▼	0.871▼	0.923▼
	48	MSE	**0.827**	0.931▼	0.878▼	1.081▼	1.062▼	1.130▼	1.518▼
		MAE	**0.777**	0.837▼	0.806▼	0.902▼	0.896▼	0.927▼	1.030▼
	72	MSE	**0.834**	0.953▼	0.879▼	1.091▼	1.091▼	1.134▼	1.705▼
		MAE	**0.785**	0.850▼	0.806▼	0.907▼	0.907▼	0.926▼	1.103▼
	96	MSE	**0.831**	0.960▼	0.875▼	1.091▼	1.087▼	1.127▼	1.836▼
		MAE	**0.785**	0.853▼	0.810▼	0.913▼	0.910▼	0.925▼	1.151▼
Solar Energy	24	MSE	**0.233**	0.245▼	0.277▼	0.389▼	0.406▼	0.269▼	2.416▼
		MAE	**0.268**	0.280▼	0.310▼	0.464▼	0.485▼	0.306▼	1.151▼
	48	MSE	**0.240**	0.248▼	0.272▼	0.269▼	0.269▼	0.284▼	2.406▼
		MAE	**0.276**	0.287▼	0.301▼	0.301▼	0.308▼	0.308▼	1.153▼
	72	MSE	**0.241**	0.258▼	0.255▼	0.286▼	0.291▼	0.279▼	2.470▼
		MAE	**0.269**	0.285▼	0.294▼	0.318▼	0.317▼	0.312▼	1.174▼
	96	MSE	**0.233**	0.253▼	0.235▼	0.310▼	0.275▼	0.271▼	2.476▼
		MAE	**0.262**	0.282▼	0.272▼	0.320▼	0.304▼	0.303▼	1.174▼

4.2 Background: Temporal attention

CNN-trans includes the temporal context into the attention mechanism to build a more robust attention model. It is accomplished by deriving query and key vectors from the temporal window of observations preceding the time step i. Denoting this temporal window $i_{<W} = [i - W, \ldots, i]$, the transformer's attention has only needs to be modified by how query and key vectors are obtained: $\boldsymbol{q}_i = \texttt{proj}(\boldsymbol{x}_{i<W})$ and $k_j = \texttt{proj}(\boldsymbol{x}_{j<W})$. It has been shown that deriving the query $\boldsymbol{q}_i$ and key $\boldsymbol{k}_j$ vectors from a temporal window of observations enables the attention mechanism to better understand the underlying temporal behavior of the query $\boldsymbol{q}_i$ and key $\boldsymbol{k}_j$ [9]. While the incorporation of the temporal context can be helpful, a limitation is that he temporal window is of fixed size W. Even after tuning the window size W, or choosing a larger temporal window issues remain: (1) noise induced by an excessively large window may hinder good performance. (2) dynamic time periods (e.g. storms) may need a shorter window than stagnant time periods (e.g. droughts). (3) some temporal patterns may be stretched to slightly different time scales, such would not be recognized by the CNN-trans approach (see Figure 1). In the following, we address this shortcoming with an adaptive approach.

4.3 Adaptive Temporal-aware Attention (Ours)

The temporal behavior of a time series can change from fast and dynamic to slow and steady. Hence, our goal is to build a robust and adaptive model that respond to distribution shifts on different time scales as needed. Furthermore, we want the model to have this ability to identify similar behaviors regardless of their time scales. For example, if the unseen data exhibits fast and dynamic behavior, the desired model should be able to identify a similar temporal pattern even if it is slightly stretched in the past observations. Considering the storm forecasting domain, storms can be short and extreme or long and steady, where both of which generally produce similar outcomes, therefore identifying and matching these temporal behaviors is critical for accurate forecasting. To achieve this goal, we

will consider multiple temporal window sizes $\mathcal{W} = [w_1, \ldots, w_n]$ to derive query and key vectors using MLP-projections:

For all temporal windows $w \in \mathcal{W}$ we obtain: query $\boldsymbol{q}_{i,w} = \texttt{proj}(\boldsymbol{x}_{i<w})$ and key $\boldsymbol{k}_{j,w} = \texttt{proj}(\boldsymbol{x}_{j<w})$. With this approach at each position i and j, there exists temporal-aware query and key vectors that represent the temporal behavior across different time scales via different window sizes. Therefore, it allows the attention mechanism to match queries with keys that represent similar temporal behaviors regardless of the pace at which they occur. This comparison enables the attention mechanism to identify the similarity on different time scales from stretched to compressed. (see Figure 1). This would result in a more robust model under distribution shifts, and hence a better forecasting model. Our experimental evaluation will demonstrate the validity of this claim.

After deriving query and key vectors for all temporal windows, a time series with Q queries and K keys would contain $|\mathcal{W}| \cdot Q$ query and $|\mathcal{W}| \cdot K$ key vectors. Note that this approach leads to redundancies, since any input $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are subsumed into multiple overlapping context windows of query $\boldsymbol{q}_{i,w}, \boldsymbol{q}_{i+1,w}, \ldots$ and of key $\boldsymbol{k}_{j,w}, \boldsymbol{k}_{j+1,w}, \ldots$ vectors. We exploit these redundancies by selecting an optimal subset of query and key vectors with total number of Q and K queries and keys. Then use query and key vectors at indexes $i \in \mathcal{I}_s$ and $j \in \mathcal{J}_s$ to calculate the attention scores $\boldsymbol{a}_{i,j} = \texttt{softmax}(\boldsymbol{q}_i^\intercal \cdot \boldsymbol{k}_j)/\sqrt{d}$, where $\mathcal{I}_s \subset \{1, 2, \ldots, |W| \cdot Q\}$ and $\mathcal{J}_s \subset \{1, 2, \ldots, |W| \cdot K\}$ denote the selected subset of query and key positions. These positions are selected to represent the top Q and K highest valued query and key vectors. The chosen subsets meet the training objective to minimize the forecasting loss. The intuition is that during training the back-propagation encourages higher weights on the subset of query and a key vectors that contribute to a better overall forecasting performance.

5 Experiments

Datasets: We empirically perform experiment on three datasetes that have been widely used for evaluations by previous studies: (1) Traffic [2] is a collection of hourly occupancy rate of 440 SF Bay Area freeways. (2) Air quality[3] is a collection of hourly air pollution of ten cities in China. (3) Solar energy[4] is a collection of hourly solar power in different locations in America. **Implementation details**: All neural models are trained and evaluated three times. We use Optuna [1] for hyper-parameter optimization. The model size is chosen from {16, 32} for all neural models, and the batch size is set to 256. **Baselines**: (1) ARIMA [3] (2) LSTM [7] (3) Transformer [15]: a multi-layer encoder-decoder with basic attention. (4) Conv-trans [9]: a encoder-decoder with convolutional attention. (5) Autoformer [6] and (6) Informer [16].

5.1 Main Results

Results are reported as mean and standard error of MSE and MAE scores. Table 1 summarizes the evaluation results on three datasets when generating predictions for horizons of 24, 48, 72, and 96. Across all datasets and settings our model outperforms other methods. When predicting for 96 horizons Our model exhibits an MSE reduction compared to the next best model by 6% (at 96 horizons), 5% (at 72 horizons), and 5% (at 72 horizons) on traffic, air quality, and solar energy datasets. The difference in performance to CNN-trans demonstrates the gain when providing the model with the flexibility to choose the right temporal context in response to distribution shifts.

6 Conclusion

In this paper, we study the multi-horizon time series forecasting problem in the presence of distribution shifts in the timescale of temporal patterns and introduce the adaptive temporal attention, which adaptively respond to such changes, and can detect patterns if temporarily stretched or compressed. Our experimental evaluation on three datasets demonstrates performance improvements over state-of-the-art methods, including temporal CNN-trans, and generalized transformer-based models such as Autoformer and Informer. Given the complementary nature of our adaptive temporal approach,

[2] https://archive.ics.uci.edu/ml/machine-learning-databases/00204

[3] https://archive.ics.uci.edu/ml/machine-learning-databases/00501/PRSA2017-Data.zip

[4] https://www.nrel.gov/grid/assets/downloads/al-pv-2006.zip

it can be integrated into future forecasting methods whenever relevant temporal patters need to be recognized across shifting timescales.

References

[1] Takuya Akiba et al. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2623–2631. ISBN: 9781450362016. DOI: 10.1145/3292500.3330701. URL: https://doi.org/10.1145/3292500.3330701.

[2] Ahmed M. Alaa and Mihaela van der Schaar. "Attentive State-Space Modeling of Disease Progression". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1d0932d7f57ce74d9d9931a2c6db8a06-Paper.pdf.

[3] G. E. P. Box and G. M. Jenkins. "Some Recent Advances in Forecasting and Control". In: *Journal of the Royal Statistical Society Series C* 17.2 (June 1968), pp. 91–109. DOI: 10.2307/2985674. URL: https://ideas.repec.org/a/bla/jorssc/v17y1968i2p91-109.html.

[4] Pascal Courty and Hao Li. "Timing of Seasonal Sales". In: *The Journal of Business* 72.4 (Oct. 1999), pp. 545–572. DOI: 10.1086/209627. URL: https://ideas.repec.org/a/ucp/jnlbus/v72y1999i4p545-72.html.

[5] Chenyou Fan et al. "Multi-Horizon Time Series Forecasting with Temporal Attention Learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2527–2535. ISBN: 9781450362016. DOI: 10.1145/3292500.3330662. URL: https://doi.org/10.1145/3292500.3330662.

[6] haixu wu haixu et al. "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 22419–22430. URL: https://proceedings.neurips.cc/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf.

[7] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[8] Yann LeCun and Yoshua Bengio. "Convolutional Networks for Images, Speech, and Time Series". In: *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258. ISBN: 0262511029.

[9] Shiyang Li et al. "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[10] Bryan Lim. "Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

[11] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "The M4 Competition: 100,000 time series and 61 forecasting methods". In: *International Journal of Forecasting* 36.1 (2020), pp. 54–74. URL: https://EconPapers.repec.org/RePEc:eee:intfor:v:36:y:2020:i:1:p:54-74.

[12] Syama Sundar Rangapuram et al. "Deep State Space Models for Time Series Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/5cf68969fb67aa6082363a6d4e6468e2-Paper.pdf.

[13] David Salinas et al. "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3 (2020), pp. 1181–1191. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301888.

[14] Shun-Yao Shih, Fan-Keng Sun, and Hung-yi Lee. "Temporal pattern attention for multivariate time series forecasting". In: *Machine Learning* 108 (Sept. 2019). DOI: 10.1007/s10994-019-05815-0.

[15] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[16] Haoyi Zhou et al. "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting". In: *AAAI*. 2021.