Self-attention with temporal prior: can we learn more from the arrow of time?

Kyung Geun Kim[1]*[†] and Byeong Tak Lee[2†]

[1]VUNO Inc., Seoul, Republic of Korea, [2]Medical AI Co., Ltd., Seoul, Republic of Korea

Many diverse phenomena in nature often inherently encode both short- and long-term temporal dependencies, which especially result from the direction of the flow of time. In this respect, we discovered experimental evidence suggesting that *interrelations* of these events are higher for closer time stamps. However, to be able for attention-based models to learn these regularities in short-term dependencies, it requires large amounts of data, which are often infeasible. This is because, while they are good at learning piece-wise temporal dependencies, attention-based models lack structures that encode biases in time series. As a resolution, we propose a simple and efficient method that enables attention layers to better encode the short-term temporal bias of these data sets by applying learnable, adaptive kernels directly to the attention matrices. We chose various prediction tasks for the experiments using Electronic Health Records (EHR) data sets since they are great examples with underlying long- and short-term temporal dependencies. Our experiments show exceptional classification results compared to best-performing models on most tasks and data sets.

KEYWORDS

self attention, time series, Electronic Health Record (EHR), Transformer, inductive bias

1 Introduction

Time series are realizations of diverse phenomena in nature with underlying dynamics of long and short-term temporal dependencies. Although the long-term and point-wise dependencies can also be critical for certain tasks, it is natural to believe that data points that are close in time are highly *interrelational* (Tonekaboni et al., 2021). One reason this might be plausible is that the direction of time is linear and asymmetric in the physical world. Particular models with such underlying philosophy and inductive bias, like Markov Chains, are useful in many situations. With these observations, we hypothesized that allowing models to directly encode this *interrelational* time dependencies while still attending to long-term time dependencies can introduce a powerful inductive bias for better performance in relevant prediction tasks.

RNNs, including their extensions, have these inductive biases already encoded (Battaglia et al., 2018). However, it is well-known that, although RNNs can theoretically account for long-term dependencies, learning long-term dependencies with gradient descent is difficult in practice (Bengio et al., 1994). Attention-based models, such as Transformers, are designed to solve these issues (Vaswani et al., 2017). While Transformers, with the proper structure, can attend to any pairs of discrete time stamps in parallel, it has no monotonic, *interrelational* bias between data points that are close in the temporal domain (Dosovitskiy et al., 2021). We hypothesized that by explicitly enabling the capacity to learn these temporal biases, Transformers could learn relevant underlying temporal regularities that are more optimal for prediction tasks of interest.

To best achieve this, we reasoned that encouraging specific structures on the attention matrices would be the most simple and efficient solution. By designing kernels inspired by various correlation functions applied to the attention matrix directly, we propose an extension to the Transformer with Self Attention with Temporal Prior (SAT-Transformer). The designed kernels are parameterized with minimal amounts of learnable parameters to enforce temporal bias to some degree. Although the main experiments use Electronic Health Records (EHR) data sets, which take the standard form of time series data sets, we believe the SAT-Transformer is generalizable to other time series data sets. The SAT-Transformer achieved significant performance gains compared to the current best-performing model on various prediction tasks across multiple data sets. In addition, we have implemented the kernels of SAT-Transformers in a way that vectorized element-wise matrix multiplication can be possible. Compared to vanilla Transformers, SAT-Transformer only requires minimal additional computation, which can be done efficiently. As a result, SAT-Transformer is far more effective than Transformers and RNNs performance-wise while being almost as efficient as Transformers.

The main contribution of this paper is that from the idea of underlying monotonic temporal bias in time series, we have designed the SAT-Transformer that can learn to exploit these regularities directly. In Section 3, we show some direct supporting evidence for our initial hypothesis in addition to performance gains in Section 4. In addition, we show multiple ablation experiments and extensions of our model to gain more insight into the performance of our model in terms of the temporal regularities of the data sets. Through our experiments, we also show that our model can be more useful, especially in situations with a limited amount of data. We believe that the results of our paper suggest a direction to explore, which is, rather than expecting neural network models to learn from scratch, exploiting temporal biases of time series can result in room for improvement.

2 Related works

2.1 Structural bias of model and generalization

Structural bias, including inductive bias of machine learning models, has been studied ever since the field of machine learning was studied. Researchers have gradually realized that for different structures of data, different kinds of structural bias must be applied to models to achieve a jump in performance. With this wisdom, researchers have developed models such as CNNs (LeCun et al., 1999), RNNs (Jordan, 1986), Graph Neural Networks (Gori et al., 2005), or Transformers (Vaswani et al., 2017) for better prediction performance on data sets such as images, time series, and natural language.

More recently, with very similar philosophies, enabling models or learning agents to generalize to a more sophisticated task or different data formats is being explored extensively. Specifically, with the right inductive bias, one can train a reinforcement learning agent to glue blocks together to build a tower (Hamrick et al., 2018) or perform better in transfer learning problems (Li et al., 2018). In some tasks, the generalization of trained models differs not only

in performance but also qualitatively (McCoy et al., 2020). These works altogether suggest that for specific types of tasks, specific structural or inductive biases are required.
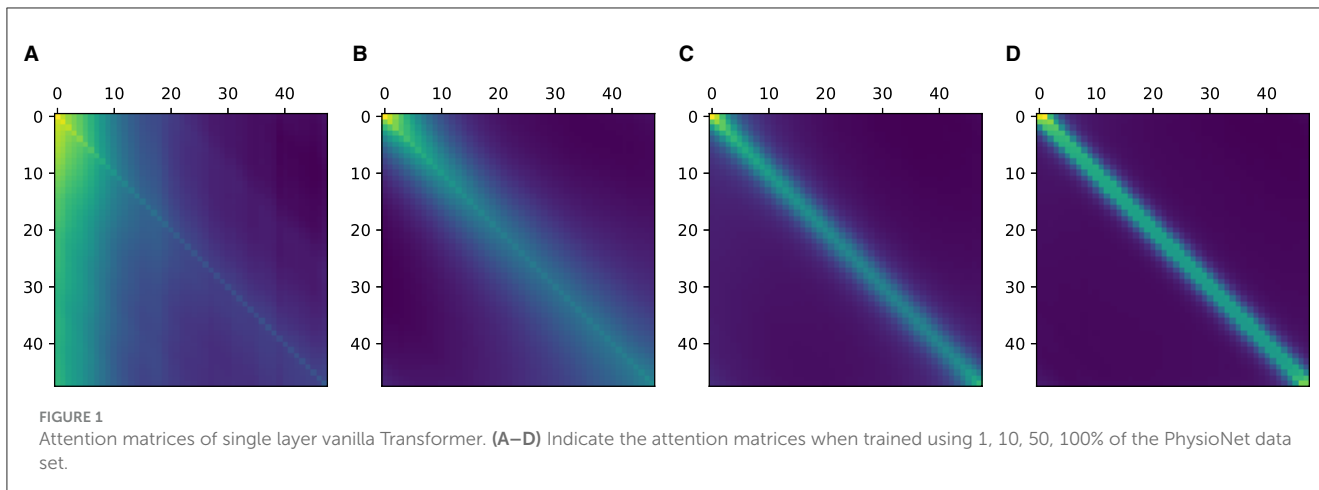
However, more evidences supporting that these structural biases might not be absolutely necessary are being discovered. A recent work done by Dosovitskiy et al. (2021) shows that Transformers can be trained to outperform the state-of-the-art CNN models on image data if the training set is extremely large. From these experiments, an intuitive conclusion can be drawn about the relation between the size of the data set and the flexibility of the model. That is, a more flexible model can eventually perform better than highly structural models when enough data are given.

2.2 Event prediction in medicine using EHR

It is widely known in medicine that prediction or detection of deterioration in the ICU can contribute to better patient management, which leads to better outcomes (Power and Harrison, 2014). While this is a critical issue, it is still a challenging task. Various machine learning-based approaches have been suggested recently to tackle these issues and discover the underlying characteristics of those patients.

Most of the works concentrate on improving the prediction performance by better imputing missing values, which have always been considered the biggest problem in EHR data analysis. The GRU-D, one of the earliest and most prominent works, predicts missing values based on the tendency that missing variables become closer to specific values as time goes by Che et al. (2018). This is done by exponential decay given to missing values and hidden states to specific values. The GRU-D outperforms vanilla GRU on the benchmark data sets by a significant margin. Phase LSTM introduces an additional time gate inside the LSTM cell. The time gate regularizes the access to the hidden state, enabling the hidden state to preserve the historical information for a more extended period (Neil et al., 2016). In another work by Li and Marlin (2016), the missing values were estimated using the Gaussian Process. The Gaussian Process and prediction network are trained together, where the output of the Gaussian Process is used as the input of the RNN. An extension of this idea is proposed in Futoma et al. (2017), which applies a multivariate Gaussian Process for imputation. Recently, Shukla and Marlin (2019) proposed an interpolation prediction network that applies multiple semi-parametric interpolation processes to obtain regularly sampled time series data. SeFT proposed by Horn et al. (2020) attempts to solve this problem from a different perspective. It directly predicts the events without the imputation process by relaxing the sequence order condition of the input data.

As the Transformers are becoming more popular in natural language processing and computer vision, many recent works utilize the Transformers to solve medically critical tasks using EHR data. In Yang et al. (2023), an encoder-decoder framework with a novel pretraining is proposed to enhance the detection of pancreatic cancer and self-harm in patients with post-traumatic stress disorder. Similarly, Meng et al. (2021) utilized bidirectional representation learning with the Transformer architecture to predict depression from EHR data. In addition, many interesting

FIGURE 1
Attention matrices of single layer vanilla Transformer. **(A–D)** Indicate the attention matrices when trained using 1, 10, 50, 100% of the PhysioNet data set.

works attempt to overcome the structural limitations of the Transformer architecture. In Li et al. (2021), the authors showed that applying a hierarchical Transformer on EHR data to expand the receptive field of the model better incorporates long-term dependencies. Because EHR data is highly irregular in time intervals, applying the Transformer directly can result in some suboptimal performance. To mitigate these issues, Peng et al. (2021) employs the Neural Ordinary Differential Equation in conjunction with the Transformer architecture, and Tipirneni and Reddy (2022) invented a novel Continuous Value Embedding technique to embed continuous time variables instead of applying discretization methods.

3 Proposed method

3.1 Learning underlying structure of a time series

We hypothesized that time series data have an underlying structure regardless of the prediction tasks. To elaborate, for a time series $\{X_t\}_{t=1}^{T}$ *interrelation* between time stamps i and j are larger for closer time stamps. Note that we do not mean this in a mathematically rigorous manner. Although this hypothesis seems rather trivial, learning relevant structures without inductive bias requires a tremendous amount of data (Dosovitskiy et al., 2021). In order to observe this behavior experimentally, we trained an extremely simple single-block transformer with a minimal amount of parameters. Though the model is too small to be useful for any real-world tasks, it is sufficient for the purpose of testing our hypothesis considering the size of the available data sets. The target task was guessing the value of a randomly chosen position on the eICU data set (Devlin et al., 2018), which is explained in detail in Section 4.1. As shown in Figure 1, we observed how the values of the attention matrix change with respect to the percentage of the data set used in training. The values along the diagonal of the attention matrix can be represented as the *interrelation* of a value at a time step to itself. We observed that the attention values converged to the diagonal as larger data samples were used. Therefore, from this experiment, we drew the following conclusions. First, as we proposed in our original hypothesis, there is indeed a higher

dependency on time stamps that are closer. Second, learning this intricate regularity requires quite a large amount of data, even for this extremely simple model.

Inspired by our original hypothesis and these elementary experiments, we devised temporal kernels for the Transformers' attention matrices. Although the experiments seem to suggest only monotonic kernels, we believe there can be additional unobserved dependencies. For this reason, the kernels should be designed with additional domain knowledge.

3.2 Architecture of the attention matrix

We first considered kernels taking the form of exponential covariance function:

$$\mathcal{C}_e\left(X_{t_1}, X_{t_2}\right) = \sigma^2 e^{-(\alpha h)^{\beta}} \tag{1}$$

where $h = |t_1 - t_2|$ for time stamps at X_{t_1} and X_{t_2} and α and β are learnable parameters. Additionally, in this paper, we considered periodic kernels because the medical data sets used in our main experiments are highly periodic:

$$\mathcal{C}_p\left(X_{t_1}, X_{t_2}\right) = \sigma^2 e^{-2\alpha^2 \sin^2(\frac{\pi h}{\beta})} \tag{2}$$

where $h = |t_1 - t_2|$ for time stamps at X_{t_1} and X_{t_2} and α and β are learnable parameters as Equation 1. Note that in Equations 1, 2, though the names of learnable parameters are the same, this is just due to the simplicity of the notation, and thus, they have different values in actual implementation. Also, we first normalize the data using z-normalization, resulting in $\sigma^2 = 1$ for simplicity of notation throughout this paper.

Let $\mathbf{A} \in \mathbb{R}^{T \times T}$ be the attention matrix with $\mathbf{V} = \{v_i\} \in \mathbb{R}^{T \times d_i}$ as the input sequence and $\mathbf{W_1}, \mathbf{W_2} \in \mathbb{R}^{d_i \times d_k}$ be the learned matrices for transforming $\mathbf{V}$ into query and key representations of $\mathbf{Q} = \mathbf{V W_1}$, $\mathbf{K} = \mathbf{V W_2}$. The kernels defined in Equations 1, 2 are applied to each element of the query and key matrices as follows:

$$\widehat{\mathbf{Q}}_{i,j} = \mathcal{C}_e\left(\mathbf{Q}_{i,i}, \mathbf{Q}_{i,j}\right) \mathbf{Q}_{i,j}$$

$$\widehat{\mathbf{K}}_{i,j} = \mathcal{C}_p\left(\mathbf{K}_{i,i}, \mathbf{K}_{i,j}\right)\mathbf{K}_{i,j}$$

getting the attention matrix with kernelization as follows:

$$\widehat{\mathbf{A}} = softmax\left(\frac{\widehat{\mathbf{Q}}\widehat{\mathbf{K}}^\top}{\sqrt{d_k}}\right)$$

However, applying these kernels element-wise is computationally inefficient because it harms parallelism. We have stacked kernel elements in a matrix format to solve this issue by applying vectorized multiplication. Since kernels are not dependent on the input's values but rather on their time stamps, a fixed kernel matrix can be derived for an attention matrix of fixed size. Therefore, the kernel matrices $\mathbf{C}^{(e)} \in \mathbb{R}^{T \times T}$ and $\mathbf{C}^{(p)} \in \mathbb{R}^{T \times T}$ can be defined as following:

$$\mathbf{C}^{(e)}_{i,j} = e^{-(\alpha|i-j|)^\beta} \tag{3}$$

$$\mathbf{C}^{(p)}_{i,j} = e^{-2\alpha^2 \sin^2(\frac{\pi|i-j|}{\beta})} \tag{4}$$

With these matrix versions of kernels, we can rewrite the attention matrix to take advantage of vectorization.

$$\widehat{\mathbf{A}} = softmax\left(\frac{(\mathbf{C}^{(e)} \odot \mathbf{Q})(\mathbf{C}^{(p)} \odot \mathbf{K})^\top}{\sqrt{d_k}}\right)$$

Where $\odot$ denotes element-wise multiplication.

4 Experiments

4.1 Data sets

We used three different open EHR data sets, which contain multiple labels described below.

The PhysioNet Challenge 2019 data set (Goldberger et al., 2000; Reyna et al., 2020) consist of hourly clinical variables collected from intensive care unit (ICU) of two hospital systems. Clinical variables include eight vital signs, 26 laboratory values, and six demographic information. The data set contains 40,336 patients with the number of rows of 1,424,171. With this data set, the task is to predict sepsis within 12 h, where the onset of sepsis is defined by sepsis-3 criteria (Singer et al., 2016; Reyna et al., 2020). Since the original data set does not contain event labels in the time window of 6–12 h prior to the onset of sepsis, we inserted additional labels to the corresponding points. The number of septic patients is 2,932, accounting for 7.2% of entire subjects.

MIMIC-III (Johnson et al., 2016) is a multivariate clinical time series database collected at Beth Israel Deaconess Medical Center. We processed the data set for mortality prediction based on the method defined at Harutyunyan et al. (2019). The extracted data set contains 21,139 ICU stays, and each ICU stay includes 17 clinical variables with measurement intervals of 1 h. The objective is to predict in-hospital mortality using the measurements from the first 48 h of ICU stay. The number of positive cases is 2,797, which is 13.2% of the total number of patients.

The eICU Collaborative Research Database (Pollard et al., 2018) is a freely available multi-center database containing over 200,000 admissions to ICU. The database was collected from 335 units at 208 hospitals in the US. We processed the data set following the procedure defined at the PhysioNet challenge 2019 (Reyna et al., 2020). Additionally, we excluded patients if there was at least one unmeasured time stamp. Among 200,859 admissions, 132,112 admissions with 7,692,965 rows are extracted as a result. Each of the data points includes 40 variables measured in one-hour intervals. Here, we defined multiple tasks of predicting Heart failure, Respiratory failure, and Kidney failure within 12 h prior to diagnosis. Each event is set based on the ICD-9 code (HF: 428.0, RF: 518.81, KF: 584.9) on EHR. The number of patients is 4,286 (3.24%), 8,279 (6.27%), and 4,357 (3.30%) for each task.

4.2 Models and settings

We compared our method to the widely used, best-performing models in time series prediction. (1) GRU-Simple: an extension of GRU that takes time series and information about missing variables as input. (2) GRU-D: an extension of GRU that implements hidden state decays and missing value decays, encouraging convergence to the mean value. (3) Interpolation Network: predicts missing values using adjacent measurement and radial basis functions. (4) SeFT: encodes each observation separately as an unordered set and pooling them together in the attention layer. (5) Transformer: uses self-attention matrices, which considers the entire sequence to encode the relationship between each time stamp.

Unlike the MIMIC-III data set, the PhysioNet and eICU data set does not contain a fixed number of sequence lengths, making the prediction to be made using varying lengths of time series. To limit the memory requirement, we used the window of 48 h prior to prediction points. For all data sets, we used 80% of the instance as a training set and the rest as a testing set. Within the 80% of the training set, 20% of the data are randomly selected as the validation set. For PhysioNet and MIMIC-III, the test sets were constructed to be exactly the same as Horn et al. (2020) for a fair comparison. The test set for eICU was created by randomly selecting 20% of the data samples.

Since EHR data sets are highly skewed, we constructed a balanced batch by sampling event and non-event cases to the same ratio during training. The area under the precision-recall curve (AUPRC) was used to measure the validation performance of the model to determine the best hyperparameters. We stopped training if AUPRC does not improve for 30 epochs. In the tuning process, more than 200 hyperparameter sets are explored for each model trained on PhysioNet and MIMIC-III data sets. In addition, we repeated the experiments 3 times using different random seeds for stable performance. The eICU database contains more than 7,000,000 rows, making it difficult to conduct extensive hyperparameter searches. Thus, we explored smaller hyperparameter space where each hyperparameter set was tested twice with different random seeds. Training was stopped if the increase of AUPRC was not observed for 5 epochs. We tested

TABLE 1 Performance of models for each data set and task.

|	PhysioNet		MIMIC-III		eICU-HF		eICU-RF		eICU-KF		ms/iter
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	
GRU-S	15.0 ± 0.5	82.5 ± 0.4	51.6 ± 0.6	85.4 ± 0.4	2.46 ± 0.11	77.0 ± 0.5	3.48 ± 0.01	71.8 ± 0.1	2.65 ± 0.01	76.8 ± 0.5	24.8
GRU-D	14.9 ± 0.6	82.8 ± 0.3	52.8 ± 0.5	86.1 ± 0.7	2.92 ± 0.05	80.1 ± 0.2	3.42 ± 0.13	72.3 ± 0.7	2.60 ± 0.16	78.3 ± 0.7	41.9
IP-Nets	15.3 ± 0.2	82.7 ± 0.1	51.8 ± 0.9	85.4 ± 0.1	2.92 ± 0.01	80.8 ± 0.1	3.30 ± 0.02	73.4 ± 0.1	2.31 ± 0.03	77.1 ± 0.1	25.8
Transformer	15.0 ± 0.2	81.4 ± 0.2	49.7 ± 0.3	84.8 ± 0.2	2.92 ± 0.08	**81.3 ± 0.1**	3.58 ± 0.06	73.5 ± 0.2	3.09 ± 0.28	80.8 ± 0.3	6.71
SeFT	13.3 ± 0.2	82.0 ± 0.1	46.2 ± 0.1	85.1 ± 0.3	**3.16 ± 0.01**	78.3 ± 0.7	3.32 ± 0.03	72.4 ± 0.3	3.08 ± 0.30	78.3 ± 0.1	3.78
SAT-Trans	**16.7 ± 0.3**	**83.0 ± 0.7**	**53.7 ± 0.5**	**86.4 ± 0.3**	3.10 ± 0.09	81.0 ± 0.3	**3.80 ± 0.03**	**73.7 ± 0.2**	**3.33 ± 0.03**	**81.0 ± 0.5**	7.23

Bold represent the highest values among them.

around 50 hyperparameter sets for each model for each task from the eICU data set. Further, all ablation studies and extension experiments are done on the PhysioNet data set. As a result, the baselines shown throughout this paper are at least as good as the baselines used in many works in literature, if not better. All the details discussed in this section, including baseline references and search space for hyperparameters, are included in the provided Supplementary material.

4.3 Results

Table 1 shows the performance of the SAT-Transformer against other well-known models with the best performance up to date. Regarding AUPRC, which is the most appropriate metric for unbalanced data sets, SAT-Transformer performs significantly better than other models in almost every task across multiple data sets. In terms of AUROC, while it is highly affected by the imbalance in the data sets, SAT-Transformer still performs best in most of the tasks. Additionally, while the second best-performing model significantly varies for different tasks and data sets, SAT-Transformer achieves that best performance significantly across all tasks. To measure computational efficiency, time (in milliseconds) per iteration has been measured and averaged over multiple trials. As a result, the SAT-Transformer is comparable to or only slightly slower than the vanilla Transformer and much more efficient than other RNN-based models.

To demonstrate the effect of the encoded inductive bias of the SAT-Transformer according to the size of the data set, we examined the drop in performance when only a subset of data is used in training. We selected GRU-Simple, vanilla Transformer as a representative in each category. As shown in Figure 2, the performance difference between SAT-Transformer and GRU-Simple becomes significant as the size of the data set increases. Another point to note is that the performance increase of the vanilla Transformer becomes more significant as more data samples are used, achieving similar performance as GRU-Simple when 100% of the data set was used. A recent study in Dosovitskiy et al. (2021) suggests that models with more flexibility or less inductive bias can eventually outperform models with more structures in the case of

FIGURE 2
Performance of each model with respect to reduction of data set size.
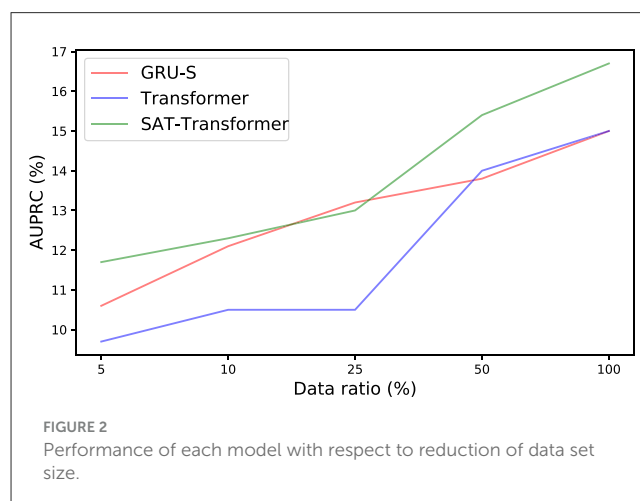
extremely large data sets. Similar behavior can be observed when the two models are compared with SAT-Transformer. Since GRU-Simple is highly structured due to the recurrent architecture when dealing with time series and Transformers are more flexible, a less dramatic drop in the performance of GRU-Simple makes sense. The SAT-Transformer, however, can be thought of as being almost as flexible as the Transformer while it is capable of maintaining the inductive bias of the RNNs. When given an extremely large data set, the vanilla Transformer may eventually outperform every model, but the performance difference shown in Table 1 suggests that this is currently almost infeasible.

4.4 Ablation study and extensions

The periodic kernels have been added due to the prior domain knowledge of our medical data sets. To see the effect of this, we performed an ablation study using only exponential kernels and only periodic kernels. The experiments are done on PhysioNet data sets with sepsis prediction as the target task. As the results summarized in Table 2 show, the model performs better than the vanilla Transformer when any kernels are used. Comparing

exponential kernels and periodic kernels, the performance gain resulting from exponential kernels is more significant than periodic kernels.

We explored two types of additional extensions that can be applied to SAT-Transformer. First, to demonstrate the compatibility of SAT-Transformer as a module with the existing framework, we added the Interpolation Network module defined in Section 3.2.1 of Shukla and Marlin (2019) on top of our model. We also trained a vanilla transformer with an Interpolation Network added to the top for comparison. These experiments are done on PhysioNet data sets with sepsis prediction as the target task. As shown in Table 3, both models with Interpolation Network gain additional performance as expected.

Second, because time series data can be highly diverse, kernels that depend on temporal features can be beneficial. For example, in certain prediction tasks involving EHR data, the length of the vital sequence can be a significant factor affecting the prediction performance (Li et al., 2019). Therefore, we designed temporal feature adaptive kernels that the learnable parameters α and β from Equations 3, 4 can be adaptive. Specifically, α and β are computed from a linear model taking temporal features for each data point $\{v_t\}_{t=1}^{T}$ as an input. The temporal feature inputs are defined as vectors of mean, standard deviation, sequence length, and average time interval ($\frac{t_T - t_1}{T}$). The experiments are done on every EHR data set as presented in Table 1, and the results are presented in Table 4.

There are some minor improvements for some tasks using adaptive kernels, but they are relatively inconsistent.

5 Discussion

From our original hypothesis, we have proposed the SAT-Transformer, verifying that it significantly outperforms other best-performing models up to date in multiple prediction tasks across various medical data sets. In this section, we further examine the learned attention matrices and kernels of SAT-Transformer from the perspective of the original hypothesis.

We first examine the learned attention matrices of the SAT-Transformer and compare them against the attention matrices of the vanilla Transformer. To show the effect across every element of the model, we present the attention matrices for all three layers for two different heads. As shown in Figure 3, many attention matrices in SAT-Transformer reveal higher *interrelation* between closer time stamps. In addition, the effects of periodic kernels are observed in some attention matrices. In contrast, in the exact same matrices in the vanilla transformer, such behaviors are unobserved, showing rather similar attention maps across multiple heads and layers. From these observations, we conclude that with the assistance of the temporal kernels, the model was able to learn a more relevant attention scheme for given medical data sets and tasks of interest. In addition, with the previous experiment shown in Figure 1 and attention matrices shown in Figure 3, we demonstrate supporting pieces of evidence for our original hypotheses.

In Figure 4, we show the shape of learned kernels without the effect of the attention for multiple heads in all three layers. Here, we observed that diverse kernels are learned for different attention heads. In many cases, the effect of exponential and periodic kernels coexist, but one of the kernels is dropped in some cases. Kernels in some heads are tuned to have almost no impact by obtaining values that are mostly close to 1 when necessary.

A recent study suggested that attention matrices in Transformers with multiple heads tend to become very similar to each other, collapsing into a single matrix (Voita et al., 2019). This work suggests that because of this lack of diversity, Transformers experience some limitations in its potential. In SAT-Transformer, due to the effect of kernels, we found that attention matrices became much more diverse compared to vanilla Transformer. Although this may not be the main reason SAT-Transformer reaches exceptional performance, we believe this might be one of the minor reasons.

TABLE 2 Performance of SAT-Transformer according to the use of different kernels.

Kernels	AUPRC	AUROC
No kernel	15.0 ± 0.2	81.4 ± 0.2
Exp	16.3 ± 0.6	$\mathbf{83.2 \pm 0.2}$
Periodic	15.9 ± 0.3	81.7 ± 0.1
Exp & periodic	$\mathbf{16.7 \pm 0.3}$	83.0 ± 0.7

Bold represent the highest values among them.

TABLE 3 Performance of Transformer models with Interpolation Network added on top.

Models	AUPRC	AUROC
Transformer	15.0 ± 0.2	81.4 ± 0.2
Transformer & IP-Net	15.4 ± 0.6	82.0 ± 0.3
SAT-Transformer	$\mathbf{16.7 \pm 0.3}$	$\mathbf{83.0 \pm 0.7}$
SAT-Trans & IP-Net	$\mathbf{17.5 \pm 0.2}$	$\mathbf{83.7 \pm 0.3}$

Bold represent the highest values among them.

TABLE 4 Effect of temporal feature adaptive kernels on performance.

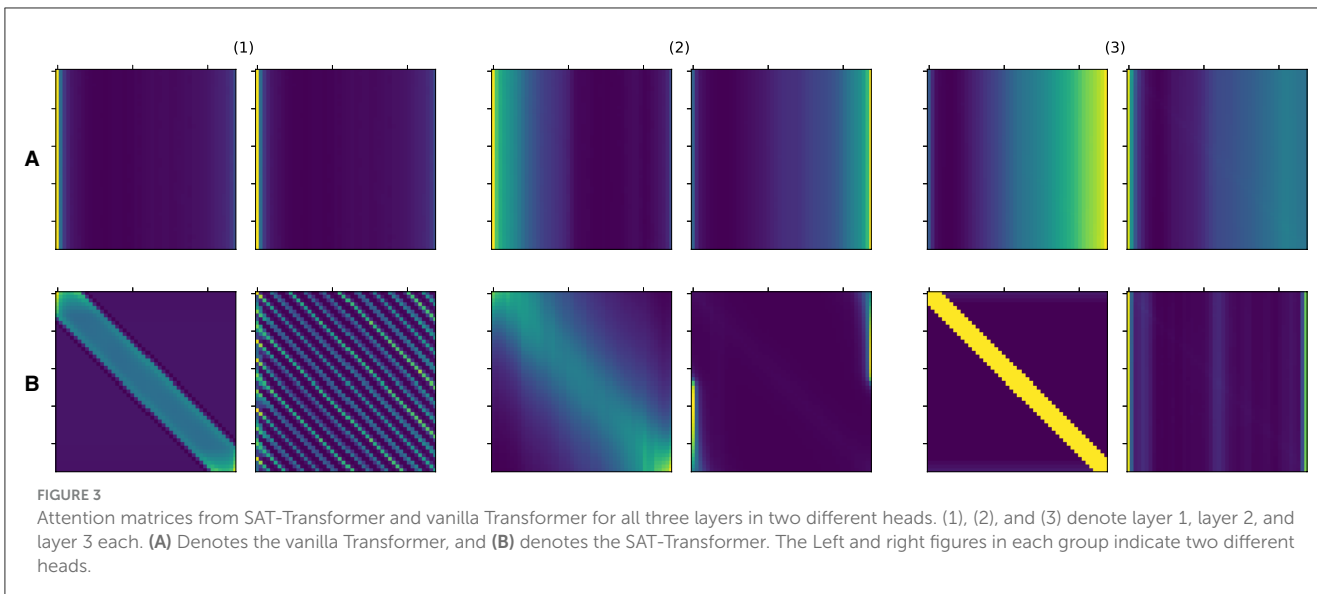|	PhysioNet		MIMIC-III		eICU-HF		eICU-RF		eICU-KF	
	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC
Non-adaptive	$\mathbf{16.7 \pm 0.3}$	83.0 ± 0.7	$\mathbf{53.7 \pm 0.5}$	$\mathbf{86.4 \pm 0.3}$	$\mathbf{3.10 \pm 0.09}$	81.0 ± 0.3	$\mathbf{3.80 \pm 0.03}$	$\mathbf{73.7 \pm 0.2}$	$\mathbf{3.33 \pm 0.03}$	$\mathbf{81.0 \pm 0.5}$
Adaptive	16.4 ± 0.3	$\mathbf{83.3 \pm 0.3}$	53.2 ± 0.6	86.0 ± 0.3	3.06 ± 0.06	$\mathbf{81.0 \pm 0.4}$	3.70 ± 0.01	72.8 ± 0.2	3.26 ± 0.01	80.7 ± 0.1

Bold represent the highest values among them.

FIGURE 3
Attention matrices from SAT-Transformer and vanilla Transformer for all three layers in two different heads. (1), (2), and (3) denote layer 1, layer 2, and layer 3 each. **(A)** Denotes the vanilla Transformer, and **(B)** denotes the SAT-Transformer. The Left and right figures in each group indicate two different heads.
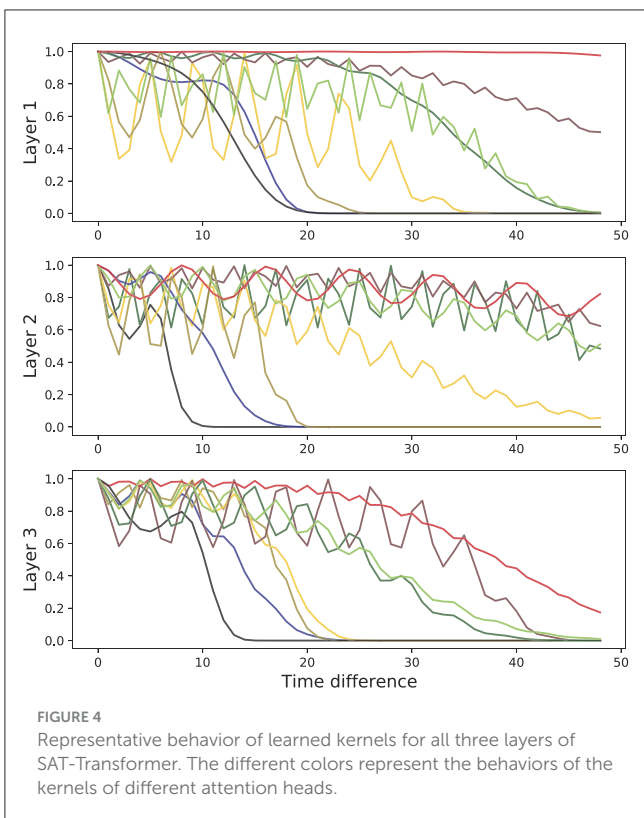

FIGURE 4
Representative behavior of learned kernels for all three layers of SAT-Transformer. The different colors represent the behaviors of the kernels of different attention heads.

that time series data arising in nature have *interrelational* bias between close data points. By taking advantage of these underlying structures, we were able to gain a significant amount of performance. Finally, although many evidences, including ours, suggest that, these regularities can eventually be learned with enough data. However, this often requires an extremely large amount of data, which is still infeasible. In this respect, imposing the right kind of inductive bias, such as the one in SAT-Transformer, can be an extremely powerful concept in the current era. In future work, we would like to apply variations of our model to multiple time series data sets, exploring more efficient and sophisticated methods to construct kernels with temporal priors.

Author's note

This research was conducted while BL was working under VUNO Inc.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

6 Conclusion

In this paper, we propose the SAT-Transformer and the elemental philosophy. We believe that our model as well as the analysis of the underlying phenomenon, provides an intuition and potential to be generalized to various time series data sets. Compared to other best-performing models known to date, SAT-Transformer has achieved outstanding gains in performance. In addition, we made an important observation of the fact

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

KK: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. BL: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing.

Funding

Conflict of interest

KK was employed by VUNO Inc. BL was employed by Medical AI Co., Ltd.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1397298/full#supplementary-material

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint* arXiv:1806.01261. doi: 10.48550/arXiv.1806.01261

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transact. Neur. Netw.* 5, 157–166. doi: 10.1109/72.279181

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-018-24271-9

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint]. doi: 10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An Image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929

Futoma, J., Hariharan, S., and Heller, K. (2017). "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proceedings of the 34th International Conference on Machine Learning (PMLR)*, 1174–1182. Available online at: https://proceedings.mlr.press/v70/futoma17a.html

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, e215–e220. doi: 10.1161/01.CIR.101.23.e215

Gori, M., Monfardini, G., and Scarselli, F. (2005). "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Vol. 2* (Montreal, QC: IEEE), 729–734. Available online at: https://ieeexplore.ieee.org/abstract/document/1555942/

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., et al. (2018). Relational inductive bias for physical construction in humans and machines. *arXiv preprint* arXiv:1806.01203. doi: 10.48550/arXiv.1806.01203

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Sci. Data* 6, 1–18. doi: 10.1038/s41597-019-0103-9

Horn, M., Moor, M., Bock, C., Rieck, B., and Borgwardt, K. (2020). "Set functions for time series," in *Proceedings of the 37th International Conference on Machine Learning (PMLR)*, 4353–4363. Available online at: https://proceedings.mlr.press/v119/horn20a.html

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.35

Jordan, M. I. (1986). Serial order: a parallel distributed processing approach. *Adv. Psychol.* 121, 471–495. doi: 10.1016/S0166-4115(97)80111-2

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision* (Berlin: Springer-Verlag), 319.

Li, S. C.-X., and Marlin, B. (2016). "A scalable end-to-end Gaussian process adapter for irregularly sampled time series classification," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (Curran Associates). Available online at: https://proceedings.neurips.cc/paper_files/paper/2016/hash/9c01802ddb981e6bcfbec0f0516b8e35-Abstract.html

Li, X., Grandvalet, Y., and Davoine, F. (2018). "Explicit inductive bias for transfer learning with convolutional networks," in *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden). Available online at: https://proceedings.mlr.press/v80/li18a.html

Li, X., Kang, Y., Jia, X., Wang, J., and Xie, G. (2019). "TASP: a time-phased model for sepsis prediction," in *2019 Computing in Cardiology (CinC)* (Singapore: IEEE), 1–4.

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., et al. (2021). Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *arXiv* [preprint]. doi: 10.48550/arXiv.2106.11360

McCoy, R. T., Frank, R., and Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Trans. Assoc. Comput. Linguist.* 8, 125–140. doi: 10.1162/tacl_a_00304

Meng, Y., Speier, W., Ong, M. K., and Arnold, C. W. (2021). Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inf.* 25, 3121–3129. doi: 10.1109/JBHI.2021.3063721

Neil, D., Pfeiffer, M., and Liu, S.-C. (2016). Phased LSTM: accelerating recurrent network training for long or event-based sequences. *arXiv* [preprint]. doi: 10.48550/arXiv.1610.09513

Peng, X., Long, G., Shen, T., Wang, S., and Jiang, J. (2021). Sequential diagnosis prediction with transformer and ontological representation. *arXiv* [preprint]. doi: 10.1109/ICDM51629.2021.00060

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The EICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* 5, 1–13. doi: 10.1038/sdata.2018.178

Power, G. S., and Harrison, D. A. (2014). Why try to predict ICU outcomes? *Curr. Opin. Crit. Care* 20, 544–549. doi: 10.1097/MCC.0000000000000136

Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Westover, M. B., Nemati, S., et al. (2020). Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Crit. Care Med.* 48, 210–217. doi: 10.1097/CCM.0000000000004145

Shukla, S. N., and Marlin, B. M. (2019). Interpolation-prediction networks for irregularly sampled time series. *arXiv* [preprint. doi: 10.48550/arXiv.1909.07782

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA* 315, 801–810. doi: 10.1001/jama.2016.0287

Tipirneni, S., and Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data* 16, 1–17. doi: 10.1145/3516367

Tonekaboni, S., Eytan, D., and Goldenberg, A. (2021). Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint* arXiv:2106.00750. doi: 10.48550/arXiv.2106.00750

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Curran Associates). Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). *Analyzing Multi-Head Self-Attention: Specialized Heads do the Heavy Lifting, the Rest Can be Pruned.*

Yang, Z., Mitra, A., Liu, W., Berlowitz, D., and Yu, H. (2023). Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.* 14:7857. doi: 10.1038/s41467-023-43715-z