An Oriented Attention Model for Infectious Disease Cases Prediction

Peisong Zhang[1][0000−0001−6990−7143], Zhijin Wang[2][0000−0002−7962−2827], Guoqing Chao[3][0000−0002−2410−650X], Yaohui Huang[4][0000−0002−3437−1120], and Jingwen Yan[5][0000−0002−6153−3519]

[1] School of Science, Jimei University,
Yinjiang Road 185, Xiamen 361021, China
pencil007123@gmail.com
[2] Computer Engineering College, Jimei University,
Yinjiang Road 185, Xiamen 361021, China
zhijinecnu@gmail.com
[3] School of Computer Science and Technology, Harbin Institute of Technology,
2 West Culture Road, Weihai, 264209, P.R. China
guoqingchao@hit.edu.cn
[4] College of Electronic Information, Guangxi University for Nationalities,
Daxue East Road 188, Nanning 530006, China,
yhhuang5212@gmail.com
[5] College of Engineering, Shantou University,
University Road 243, Shantou 515063, China
jwyan@stu.edu.cn

Abstract. Effective infectious disease prediction supports the success of infection prevention and control. Several attention-based predictive models can be applied to undertake the prediction task. However, using a single attention mechanism can only capture local information, i.e. part of the temporal dynamics from time series. In this paper, we take for the hypothesis that using multiple attention from different aspects could improve prediction accuracy. An oriented attention model (OAM) is proposed to draw temporal dynamics in several aspects, via oriented attention units and their aggregation. Firstly, time series are represented as oriented transformations. And then those representations are consolidated to connect with outputs. Intensive experiments on two real infectious disease datasets show OAM's effectiveness.

Keywords: Infectious disease · Prediction · Oriented attention · Aggregation · Time series

1 Introduction

Infectious disease is always a health problem for human beings. For example, about one million children are newly infected by hand, foot, and mouth disease (HFMD), and more than one million people are newly infected by hepatitis beta virus (HBV) every year in China [1]. The early warning system is conducive to managing the risk of infectious disease [18]. The predictive technique is the most critical part of this system [17], which supports the decision-making in healthcare and intervention strategies [10,13].

The infection cases prediction is commonly regarded as a time series prediction problem. Abundant methods have been used to predict different kinds of epidemics [7]. Owning to the success of attention mechanism [12] in time series prediction [9], it had been introduced to predicting infectious disease [23]. These attention mechanisms provide abilities in focusing on some important factors among different time intervals. In reality, temporal dynamics of infectious disease time series are usually complicated and changeable [15]. Hence, a single attention mechanism is commonly captures incomplete temporal dynamics. Meanwhile, most of these attention-based time series prediction methods are based on a single attention mechanism.

We take for the hypothesis that using multiple attention from different aspects could improve prediction accuracy. The goal is to investigate an attention representation and fusion model, which provides the ability to capture temporal dynamics of infectious disease time series from several aspects. There are two challenges as follows: (1) how to represent inputted time series in different aspects by using several attentions? (2) how to fuse those represented attentions?

An oriented attention model (OAM) is proposed to overcome the two challenges. Firstly, the inputted time series is normalized and split into a data cube of look-back windows. Secondly, an oriented attention unit (OAU) is exploited to represent the cube in four orientations. The OAU is designed to highlight information from past observations in several aspects. Finally, the oriented attentions are fused to connect with upcoming values. To evaluate the effectiveness of the proposed OAM, intensive experiments were conducted on HFMD and HBV datasets.

The main contributions of this paper are summarized as follows:

(1) The attention mechanism on time series can be represented in four aspects, other than one aspect before.
(2) Moreover, the feasibility of the fusion of different attentions has been validated.
(3) The proposed OAM significantly outperforms the state-of-the-art methods on two real datasets.

The rest of this paper is organized as follows. Section 2 reviews several relevant work. Section 3 defines the research problem. Section 4 illustrates the proposed OAM. Section 5 presents experiments and analyses. Finally, a conclusion is given in Section 6.

2 Related Work

The relevant researches are addressed at attention-based time series prediction methods [6,9] and temporal fusion methods [13,14,16,22].

Attention-based time series prediction methods. These methods added attention components in the time series representation processes. These attentions provided the ability to strengthen or weaken observations within a time interval. These attentions have been reviewed in [4]. The major part of the attention mechanism is its score function, and they are divided into multiplicative function, additive function, and multiple layer perceptron (MLP). Besides, the multiplicative function consists of dot function, scaled dot function, and general form.

The attention mechanism was first proposed in [12]. The attention-based time series prediction was first applied to stock price prediction in [9], which captured the attentions in recurrent neural networks (RNN) from both NASDAQ index time series and the top 80 stock prices time series of NASDAQ. The stacked attentions were added after convolution neural networks (CNN) to represent temporal features in multivariate time series forecasting [6]. In summary, these attentions are used to highlight the observations of time intervals from inputted time series. However, this research believes the attentions are not only limited to the dimensionality of time steps. Technically, the dimensionality of observations should be inputted to attentions as well.

Temporal fusion methods. The temporal fusion layer is used to connect represented inputs with target values, a.k.a., model outputs. The temporal fusion problem can be regarded as the problem of mapping several input values to a target value. Methods in this category can be divided into additive operation [**?**], linear mapping [15] and non-linear mapping [22].

To the best of our knowledge, the attention fusion from different aspects had not been considered in previous research. This research fuses attentions into a temporal fusion layer to learn the temporal dynamics and make predictions.

3 Problem Definition

Time series. A time series is used to denote an ordered observed sequence of outpatient cases. Let K be the length of a time series. The weekly infectious outpatient cases are denoted by $\boldsymbol{Z}$, where $\boldsymbol{Z} \in \mathbb{R}^{K\times 1}$.

Look-back window. A look-back window is used to describe observations in several consecutive time intervals. Let T be the window size. A look-back window is denoted by symbol $\boldsymbol{Z}_{t+1:t+T,1} \in \mathbb{R}^{T\times 1}$.

Time series prediction. Commonly, the time series prediction problem is formulated as:

$$\hat{Y}_{T+1,1} = F(\boldsymbol{Z}_{t+1:t+T,1}), \tag{1}$$

where $\hat{Y}_{T+1,1} \in \mathbb{R}$ is the model output, $\boldsymbol{Z}_{t+1:t+T,1}$ is the model input, and $F(\cdot)$ is a mapping.

Attention-based time series prediction. The problem of attention-based time series prediction problem is formulated as:

$$\hat{Y}_{T+1,1} = F(A(\boldsymbol{Z}_{t+1:t+T,1})), \tag{2}$$

where $A(\cdot)$ is an attention mechanism. It should be noted that the model inputs may be processed using neural networks before feeding into the attention component.

Oriented attention-based time series prediction. The problem of oriented attention-based time series prediction problem is formulated as:

$$\hat{Y}_{T+1,1} = F(A^1(\boldsymbol{Z}_{t+1:t+T,1}), A^2(\boldsymbol{Z}_{t+1:t+T,1}), \cdots), \tag{3}$$

where $A^i(\cdot)$ is the oriented attention mechanism. It tells the oriented attention on the window size dimension and the time series dimension. The length of the time series dimension is 1.

The two dimensions for oriented attentions may be so limited. Hence utilizing attention on the dimension of predictive values is taken into consideration. Let B be the number of time consecutive batched inputs. Let $\boldsymbol{X} \in \mathbb{R}^{B\times T\times 1}$ be the time consecutive batched model inputs, and let $\boldsymbol{Y} \in \mathbb{R}^{B\times 1}$ be the outputs. Therefore, Equation 3 is re-formulated as:

$$\hat{\boldsymbol{Y}} = F(A^1(\boldsymbol{X}), A^2(\boldsymbol{X}), \cdots). \tag{4}$$

According to the three dimensions, the oriented attentions are extended in several aspects. The main symbols are listed in Table 1.

Table 1. Symbols and meanings.

Symbol	Semantic
K	time step number
T	look-back window size
$\boldsymbol{Z}$	the time series of outpatient cases, $\boldsymbol{Z} \in \mathbb{R}^{K\times 1}$
B	the number of time consecutive batched inputs
$\boldsymbol{X}$	batched input tensor $\boldsymbol{X} \in \mathbb{R}^{B\times T\times 1}$
$\boldsymbol{Y}$	batched output tensor $\boldsymbol{Y} \in \mathbb{R}^{B\times T\times 1}$
$F(\cdot)$	A mapping
$A(\cdot), A^i(\cdot)$	attention and oriented attention, respectively
S	The number of aspects for oriented attention
$\boldsymbol{M}^l$	The left mapped tensor of OAU
$\boldsymbol{M}^n$	The normalized tensor of OAU
$\boldsymbol{M}^r$	The right mapped tensor OAU
$\boldsymbol{P}$	Concatenated representation

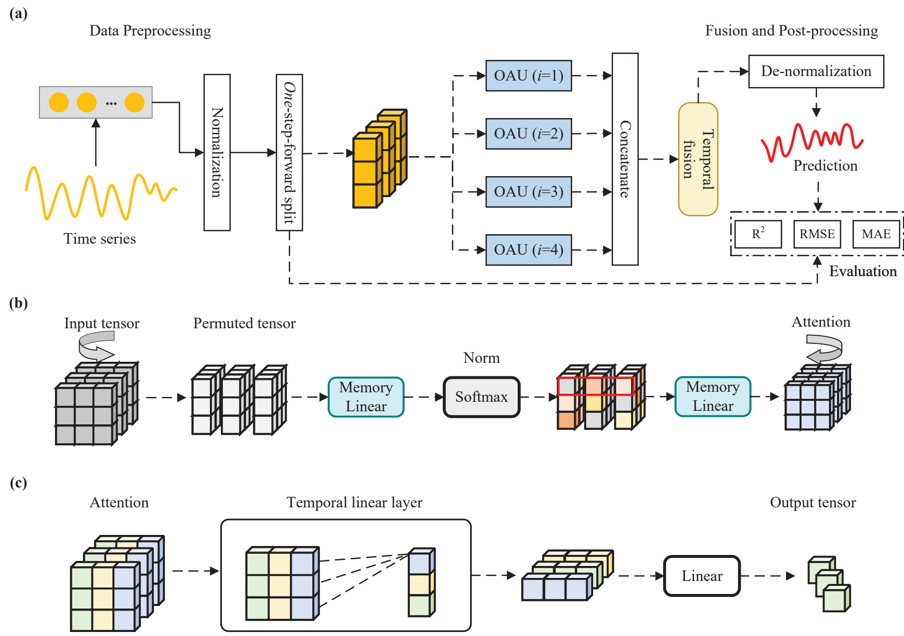


Fig. 1. The schematic illustration of the proposed oriented attention model (OAM). (a) The workflow. (b) Orientable attention unit (OAU). (c) Temporal fusion layer.

4 The Proposed OAM

The schematic illustration of the proposed OAM is displayed in Figure 1. According to the workflow in subfigure 1(a), this model consists of three stages. The first stage is data processing. The inputted time series are normalized using Min-Max normalization [17]. There are some reasons for the normalization, such as computational speed and the computability of some neural networks. The normalized time series are transformed to supervised data using "*one*-step-forward split" [18]. The second stage is the generation of oriented attentions and their combinations. Those combinations are fed into a temporal attention layer, and model outputs are de-normalized [17] in the third stage.

4.1 Oriented attention unit (OAU)

The detailed process of the oriented attention unit (OAU) is plotted in Figure 1(b). To capture the attention from several aspects and reduce the model complexity, the aspect selection component and the attention representation component are designed in OAU.

There are many ways to view the input tensor in different aspects. A simple way is to transpose the input tensor, which guarantees the attention mechanism can work well with two dimensions of it.

The *aspect selection component* views the input tensor by transposing its dimensions. Since the batch dimension may change, when the model optimizer changes. Hence, the batch dimension can not be put in the feature dimension. For the three dimensions of an input tensor, we have and only have four aspects.

Given an input tensor $\boldsymbol{X} \in \mathbb{R}^{B\times T\times 1}$, the four aspects of on it can be formulated as follows:

$$\boldsymbol{X}^i = \begin{cases} \boldsymbol{X}, & i=1 \\ \boldsymbol{X}.permute(0,2,1), & i=2 \\ \boldsymbol{X}.permute(2,0,1), & i=3 \\ \boldsymbol{X}.permute(1,0,2), & i=4 \end{cases} \tag{5}$$

where $\boldsymbol{X}^1 \in \mathbb{R}^{B\times T\times 1}$, $\boldsymbol{X}^2 \in \mathbb{R}^{B\times 1\times T}$, $\boldsymbol{X}^3 \in \mathbb{R}^{1\times B\times T}$, and $\boldsymbol{X}^4 \in \mathbb{R}^{T\times B\times 1}$ are the four kinds of aspects, respectively. The $permute(\cdot)$ function operates views on the input tensor by transposing dimensions.

The oriented attention mechanism works on the last two dimensions of an input tensor. For example, $\boldsymbol{X}^1$ is adopted to distinguish the impact of different time intervals by mapping values within a time interval. Technically, $\boldsymbol{X}^2$ is exploited to distinguish the impact of different time series by mapping the lock-back window of a time series. In this paper, only one target time series is considered. Especially, $\boldsymbol{X}^3$ is used to distinguish the impact of different instances of a batch by mapping values within a time interval. $\boldsymbol{X}^4$ is used to distinguish the impact of different instances of a batch by comparing different time series.

The *attention representation component* consists of a symmetrical structure of two linear layers and normalized layers among them. To guarantee that the attention mechanism has a strong ability in learning different aspects and the fusion of multiple oriented attentions, a symmetrical structure is employed. The detailed structure is drawn in subfigure 1(b).

Since OAU works with the last two dimensions of the input tensor, one aspect of the attention can be an example for other aspects. Taking the attention on X^2 for example, it is first mapped using the batch matrix multiplication. The mapping process are formulated below:

$$\boldsymbol{M}^l = \boldsymbol{X}^2 \cdot \boldsymbol{W}^l, \tag{6}$$

where $\boldsymbol{M}^l \in \mathbb{R}^{B\times 1\times T}$ is the left mapped tensor, and $\boldsymbol{W}^l \in \mathbb{R}^{T\times T}$ is the linear weights.

And then a normalized layers is used to enlarge the differences of look-back windows. The normalized layer adopts the softmax operation and formulated as below:

$$M^n_{b,i,t} = \frac{\exp(M^l_{b,i,t})}{\sum_{t=1}^{T} \exp(M^l_{b,i,t})}, \tag{7}$$

where $\boldsymbol{M}^n \in \mathbb{R}^{B\times 1\times T}$ is the normalized tensor. The reason of using softmax to normalize is that, it can enlarge the differences of a target dimension by considering its attributions.

The normalized tensor is mapped to generate attentions. It's formulated as below:

$$\boldsymbol{M}^r = \boldsymbol{M}^n \cdot \boldsymbol{W}^r, \tag{8}$$

where $\boldsymbol{M}^r \in \mathbb{R}^{B\times 1\times T}$ is the right mapped tensor, and $\boldsymbol{W}^r \in \mathbb{R}^{T\times T}$ is the linear weights. The attention $\boldsymbol{A}^2 \in \mathbb{R}^{B\times 1\times T}$ on aspect $s = 2$ is obtained by transposing the dimensions in $\boldsymbol{M}^r$.

The oriented attention of four aspects are obtained using Equation 6-8. They can be denoted by symbols $\boldsymbol{A}^1, \boldsymbol{A}^2, \boldsymbol{A}^3, \boldsymbol{A}^4$.

4.2 Temporal fusion layer

The scheme of the temporal fusion layer is plotted in Figure 1(c). The temporal fusion layer is used to aggregate several oriented attention, as well as inputs.

The four oriented attention and the input are concatenated as follows:

$$\boldsymbol{P} = [\boldsymbol{A}^1; \boldsymbol{A}^2; \boldsymbol{A}^3; \boldsymbol{A}^4; \boldsymbol{X}], \tag{9}$$

where $\boldsymbol{P} \in \mathbb{R}^{B\times T\times 5}$ is the concatenated representations, and $[;]$ denotes the concatenation operation.

Several existing methods can map the representations to predictions $\boldsymbol{O} \in \mathbb{R}^{B\times 1}$, such as global auto-regression (GAR), auto-regression (AR), and multiple linear regression (MLR) [15]. GAR is adopted to do the mapping since it's simple. The predictions $\boldsymbol{O}$ are de-normalized to compare with real values while training.

5 Experiments

5.1 Settings

Datasets. 49677 and 48359 real-world HFMD and HB outpatient records were collected and used to evaluate the proposed OAM and baseline methods. In the data collecting stage, the HB outpatient cases were reported when the transaminase exceeds the twice standard. The basic statistics of the two datasets are shown in Table 2. Consequently, there are certain errors in the outpatient case data of HB. The real-world HFMD and HB outpatient cases are shared by the Xiamen Center for Disease Control and Prevention (XCDC). The outpatient cases are divided into two parts, the first part from January 5, 2015 to December 23, 2019 is used to train the models; the remains were used to validate those trained models.

Baselines. To study the benefits of attentions, several models have been developed and applied on the two real datasets. To study the benefits of oriented attentions, these models have been extended with self-attentions. Hence, those benefits can be observed by measuring the prediction performance of models.

The comparable models are listed as follows:

(1) Auto-Regression (AR) [8] generated predictions by mapping linear relationships between past observations and coming values.

Table 2. The basic description on HFMD and HB datasets. "STD" denotes standard deviation.

Dataset	Training size	Test size	Maximum	Average	Minimum	STD
Weekly HFMD cases	259	53	869	159.22	0	159.84
Weekly HB cases	259	53	282	154.99	14	38.98

(2) Long- and Short-Term Memory (LSTM) [5] exploited three gate units to capture the long-term dependency.
(3) Gated Recurrent Unit (GRU) [17] merged the hidden states and cell states of LSTM to reduce parameters.
(4) Encoder-Decoder (ED) [2] consisted of two RNN components in the encoder stage and decoder stage, respectively.
(5) Convolutional Neural Network (CNN) [3] used a CNN layer to extract temporal patterns and a linear layer to connect with outputs.
(6) CNNRNN [20] extracted the local sequential patterns to generate predictions, by leveraging a connected CNN and RNN network.
(7) Self-attention [11] is integrated into the above methods. And then we have LSTM-attn, GRU-attn, ED-attn, CNN-attn and CNNRNN-attn.

Model configurations. For a fair competition, all the training-related constant parameters are set to the same values on a dataset. To reduce the complexity in presenting the experimental results, some training-related and data-related parameters are fixed according to cross-validations on parameters. For all experiments, the learning rate is set to 0.0015.

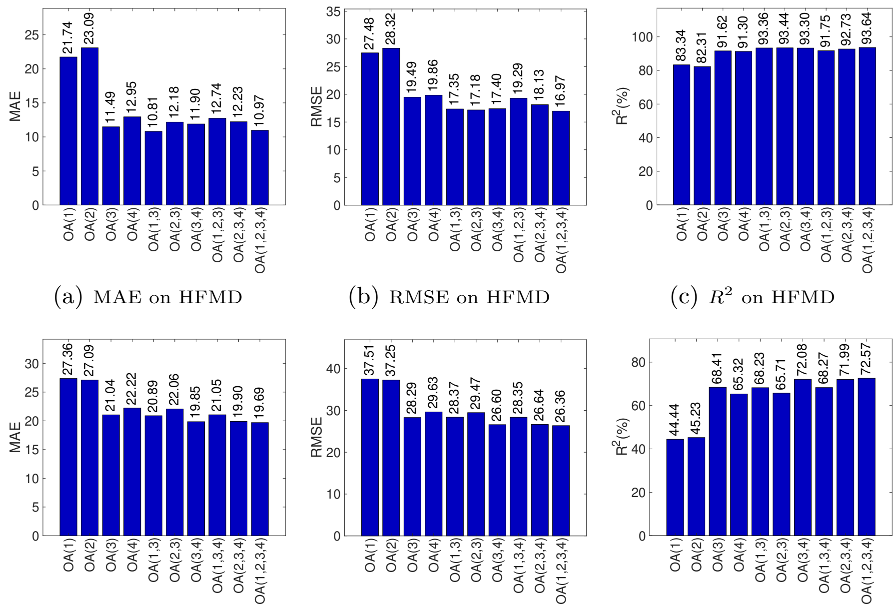
For HFMD and HB, according to the experiments on all the comparable methods, the optimal values of batch size B and look-back window size T are found at 4 and 2, respectively. Hence, $B = 4$ and $T = 2$ are both fixed in all the comparable experiments. For all the RNN-involved models, their hidden neuron sizes are both fixed at $\{32, 64\}$ to observe their prediction performance.

Performance metrics. We follow the metrics in [17,19,21]. Those metrics are mean absolute error (MAE), root mean square error ($RMSE$), and correlation coefficient (R^2). For MAE and RMSE, the lower value has better performance. For R^2, the higher value has better performance.

5.2 Study on attention combinations

To study how the oriented attentions affect the prediction performance, the experiments on OAU combinations are conducted. The experimental results on their combinations are plotted in Figure 2.

There are many situations of OAU combinations. To avoid searching for all situations, the greedy search is applied. Every aspect is done to get their corresponding performance, and the optimal aspect are selected to combined with every other aspects. Iteratively, the current optimal combination is selected to add every other aspects, until all the aspects are included.



(a) MAE on HFMD (b) RMSE on HFMD (c) R^2 on HFMD

(d) MAE on HB (e) RMSE on HB (f) R^2 on HB

Fig. 2. Comparisons of OAU combinations in terms of three metrics on two datasets.

For easy presentation, let symbol $\mathrm{OA}(i)$ be the observation in aspect set $\{i\}$. For example, the optimal single aspect attention on the HFMD dataset is found at $\mathrm{OA}(3)$, and the experiments are done to measure what aspect works best with aspect $i = 3$. This process is recursive to include all four aspects, i.e., $\mathrm{OA}(1, 2, 3, 4)$.

The major observations from Figure 2 are listed as below:

(1) For the HFMD dataset, see subfigures 2(a)-2(c), the $\mathrm{OA}(1, 2, 3, 4)$ has the optimal RMSE and R^2 values, and $\mathrm{OA}(1, 3)$ achieves the optimal MAE value.
(2) For the HB dataset, see subfigures 2(d)-2(f), the $\mathrm{OA}(3, 4)$ achieves the best performance in terms of the three metrics. The performance degrades when adding more OAUs.
(3) When observing at single OAU, $\mathrm{OA}(3)$ achieves the best performance, and $\mathrm{OA}(4)$ gets the second-best performance.
(4) $\mathrm{OA}(3)$ and $\mathrm{OA}(4)$ play the two important roles in reduce the prediction errors.
(5) $\mathrm{OA}(1)$ and $\mathrm{OA}(2)$ have a slight effects on improving the performance.

For the two datasets, the $\mathrm{OA}(1, 2, 3, 4)$ obtains optimal values in terms of RMSE and R^2. These results demonstrate that the fusion of attention mechanisms is effective. Whereas the $\mathrm{OA}(1, 3)$ achieves the optimal MAE value on the HFMD dataset, the potential reason is that the HFMD and HB dataset both are uni-variate time series. Hence the effectiveness of the attention mechanism on observation in aspects $\{2, 4\}$ becomes weak.

As shown in Figure 2, OA(3) or OA(4) significantly outperform OA(1) or OA(2). The possible reason is that OA(3) and OA(4) both can distinguish the impact of different instances of a batch and highlight the key instances among them, which assist the model capture the trends of time dynamics.

5.3 Performance comparisons

Table 3. Comparable results of twenty-two methods on the two datasets in terms of three metrics.

Model	HFMD			HB		
	MAE	RMSE	R^2	MAE	RMSE	R^2
AR	22.6962	28.0246	0.8268	27.0550	37.1552	0.4550
LSTM-32	15.8524	25.6851	0.8545	26.1595	37.6786	0.4396
LSTM-attn-32	14.2515	24.5761	0.8668	23.3605	32.0211	0.5952
LSTM-64	15.9730	25.6855	0.8545	26.6344	37.7522	0.4374
LSTM-attn-64	13.5908	24.3641	0.8691	23.2384	32.0197	0.5953
GRU-32	19.0589	26.3326	0.8471	26.1140	36.8747	0.4632
GRU-attn-32	13.6181	24.3916	0.8688	22.7751	31.1019	0.6181
GRU-64	19.5851	26.5899	0.8441	26.1305	36.9358	0.4615
GRU-attn-64	13.2111	24.2959	0.8698	22.8632	31.1539	0.6169
ED-32	18.5562	26.2696	0.8478	26.2258	37.3032	0.4507
ED-attn-32	15.1897	25.4855	0.8568	23.1548	31.4827	0.6087
ED-64	17.4964	25.8206	0.8530	25.6824	37.0903	0.4569
ED-attn-64	14.8598	25.4567	0.8571	22.6006	30.6012	0.6300
CNN-32	20.9972	27.1071	0.8379	26.4590	37.2826	0.4513
CNN-attn-32	16.1284	25.9156	0.8519	26.0703	37.0805	0.4572
CNN-64	21.0921	27.4980	0.8332	26.0738	37.3023	0.4507
CNN-attn-64	16.6176	25.8872	0.8522	26.0183	37.0168	0.4591
CNNRNN-32	19.5186	26.6080	0.8439	27.1456	36.5898	0.4715
CNNRNN-attn-32	16.1384	25.8667	0.8524	25.8596	37.1470	0.4553
CNNRNN-64	18.7608	26.3040	0.8474	26.1287	36.4199	0.4764
CNNRNN-attn-64	15.4331	25.6848	0.8545	26.0657	37.2316	0.4528
OAM	**10.9700**	**16.9735**	**0.9364**	**21.7509**	**28.9388**	**0.6692**

The performance comparison is done to validate the effectiveness of OAM. The comparable results of twenty-two methods are shown in Table 3. For all the methods, parameter batch size B is fixed at 4. The parameter look-back window size T is set to 2 and 4 for the HFMD dataset and the HB dataset, respectively.

The main results are summarized as follows:

(1) OAM achieves the best performance in terms of three metrics on the two datasets. The improvement is significant.
(2) Self-attention improves the performance for all the benchmark methods.
(3) The number of hidden neurons in RNN-involved models does not affect the performance.
(4) For the HFMD dataset, GRU-attn-64 has the second-best performance.
(5) For the HB dataset, ED-attn-64 has the second-best performance.
(6) The AR has the worst performance on the two datasets.

According to the results on the HFMD dataset in Table 3, AR had the worst prediction performance, which indicated that the linear-based methods are unable to completely fit the transmission patterns of diseases. The self-attention significantly improves the predictive accuracy of all comparable methods. GRU-attn-64 has the second-best performance, but the hidden neurons slightly affect the results. A possible reason is that increasing the hidden size of the RNN-involved models can obtain a richer representation, but the accuracy is limited by the increment of the information.

According to results on the HB dataset in Table 3, the performance of CNN-involved methods was degrading when self-attention was added. A possible reason was that HB patterns are more complicated than HFMD in terms of transmissions. CNN leads to information loss, which reduces the predictive accuracy. The performance of RNN-involved methods worked not well on the HB dataset.

The proposed OAM extracts the temporal dynamics from inputted time series by aggregating several oriented attention. Compare to other baseline methods, the proposed OAM achieved significant improvements. The MAE and RMSE values were decreased by 51.67% and 39.43% at most, respectively. The R^2 is increasing by 52.99% at most. This revealed that the fusion of oriented attention is feasible.

6 Conclusions

This paper proposed the oriented attention model (OAM) to predict the number of outpatient cases in the upcoming week, via fusing several oriented attentions. Intensive experiments on two real-world HFMD and HB data show the effectiveness of the proposed method. The attention mechanism had improved the prediction performance of several traditional methods. Moreover, the combination of oriented attention improved the performance of the single attention mechanism. This also reveals the feasibility of attention fusion and its essential in time series prediction.

In the future, the oriented attention mechanism will be further applied to multivariate time series prediction. The normalization methods, multi-instance learning will be further discussed with the attention mechanism as well.

References

1. Overview of the national epidemic situation of notifiable infectious diseases in 2020. http://www.nhc.gov.cn/jkj/s3578/202103/f1a448b7df7d4760976fea6d55834966.shtml (2022), (accessed Jan 2022)
2. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. In: Proceedings of 8th Workshop on Syntax, Semantics and Structure in Statistical Translation. pp. 103–111 (2014)
3. Hoseinzade, E., Haratizadeh, S.: Cnnpred: Cnn-based stock market prediction using a diverse set of variables. Expert Syst. Appl. **129**, 273–285 (2019)
4. Hu, D.: An introductory survey on attention mechanisms in NLP problems. In: INTELLISYS'19. vol. 1038, pp. 432–448 (Sep 2019)

5. Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., Zhang, H.: Deep learning with long short-term memory for time series prediction. IEEE Commun. Mag. **57**(6), 114–119 (2019)
6. Huang, S., Wang, D., Wu, X., Tang, A.: Dsanet: Dual self-attention network for multivariate time series forecasting. In: CIKM'19. pp. 2129–2132 (Nov 2019)
7. Keddy, K.H., Saha, S., Kariuki, S., Kalule, J.B., Qamar, F.N., Haq, Z., Okeke, I.N.: Using big data and mobile health to manage diarrhoeal disease in children in low-income and middle-income countries: societal barriers and ethical implications. Lancet Infect Dis. (Dec 2021)
8. Mabrouk, A.B., Abdallah, N.B., Dhifaoui, Z.: Wavelet decomposition and autoregressive model for time series prediction. Appl. Math. Comput. **199**(1), 334–340 (2008)
9. Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., Cottrell, G.W.: A dual-stage attention-based recurrent neural network for time series prediction. In: IJCAI'17. pp. 2627–2633 (2017)
10. Shah, W., Aleem, M., Iqbal, M.A., Islam, M.A., Ahmed, U., Srivastava, G., Lin, J.L.C.W.: A machine-learning-based system for prediction of cardiovascular and chronic respiratory diseases. Journal of Healthcare Engineering (Nov 2021)
11. Shih, S., Sun, F., Lee, H.: Temporal pattern attention for multivariate time series forecasting. Machine Learning **108**(8-9), 1421–1441 (2019)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017)
13. Wang, Y., Gu, J., Zhou, Z., Wang, Z.: Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion. Knowledge-Based Systems **88**, 12 – 23 (2015)
14. Wang, Y., Li, J., Gu, J., Zhou, Z., Wang, Z.: Artificial neural networks for infectious diarrhea prediction using meteorological factors in shanghai (china). Applied Soft Computing **35**, 280 – 290 (2015)
15. Wang, Z., Cai, B.: COVID-19 cases prediction in multiple areas via shapelet learning. Applied Intelligence **52**, 595 — 606 (2022)
16. Wang, Z., Huang, Y., Cai, B., Ma, R., Wang, Z.: Stock turnover prediction using search engine data. Journal of Circuits, Systems and Computers **30**(7), 2150122:1 – 2150122:18 (2021)
17. Wang, Z., Huang, Y., He, B.: Dual-grained representation for hand, foot, and mouth disease prediction within public health cyber-physical systems. Software: Practice and Experience **51**, 2290 – 2305 (Oct 2021)
18. Wang, Z., Huang, Y., He, B., Luo, T., Wang, Y., Fu, Y.: Short-term infectious diarrhea prediction using weather and search data in xiamen, china. Scientific Programming **2020**, 8814222:1 – 8814222:12 (2020)
19. Wang, Z., Huang, Y., He, B., Luo, T., Wang, Y., Lin, Y.: TDDF: HFMD outpatients prediction based on time series decomposition and heterogenous data fusion in Xiamen, China. In: ADMA'19. pp. 658–667. Dalian, China (Nov 2019)
20. Wang, Z., Su, Q., Chao, G., Cai, B., Huang, Y., Fu, Y.: A multi-view time series model for share turnover prediction. Applied Intelligence **In press** (2022)
21. Wang, Z., Wang, Z., Lin, Y., Liu, J., Fu, Y., Zhang, P., Cai, B.: Prediction of HFMD cases by leveraging time series decomposition and local fusion. Wireless Communications and Mobile Computing **2021**, 5514743:1–5514743:10 (2021)
22. Wu, Y., Yang, Y., Nishiura, H., Saitoh, M.: Deep learning for epidemiological predictions. In: SIGIR'18. pp. 1085–1088 (Jun 2018)

23. Zhu, X., Fu, B., Yang, Y., Ma, Y., Hao, J., Chen, S., Liu, S., Li, T., Liu, S., Guo, W., Liao, Z.: Attention-based recurrent neural network for influenza epidemic prediction. BMC Bioinform. **20-S**(18), 575:1–575:10 (2019)