**Investigating recurrent neural networks abilities in predicting influenza trends**

Tsai, Wei-Chuan

williamtsai726@gmail.com

**Abstract:**

Influenza, or flu, is a highly contagious respiratory illness which can result in hospitalization and death; thus precise and opportune prediction of the influenza trend in various region throughout the year is essential for a nation’s government to maintain public health. With accurate prediction of influenza trend in hand, the central government would be able to arrange the allocation of medical resources-apparatus, medicine, medical staff caused by the influenza. This study applies three deep neural networks, recurrent neural network(RNN) and long-short term memory(LSTM) and gate recurrent unit (GRU), to predict the trend of influenza at both regional scale and national scale. Using a real dataset from Centers for Diseases Control and Prevention, we tested the models on 10 region in the United State. The results indicate that the LSTM baseline model outperforms the other two models. Moreover, this study shows that deep learning neural network not only can be used in other fields of studies but also can be implemented as an approach to perform influenza prediction. Even though the models built in this study are only fundamental base model, we can still see the great potential deep learning neural network has in offering external assistances in influenza control and public health prevention.

**Keywords:**

Influenza prediction, deep learning, neural network, LSTM, RNN, GRU

**Introduction:**

Machine learning has underwent rapid development in the past decades. In the twenty first century, machine learning has become a popular area of study and is now widely implemented in real life situation. Natural language processing, voice recognition, hand-written digits recognition, image classification, computer vision, search engine, and virtual assistant, the application of machine learning has facilitated people’s daily life. Various neural network has been proposed for regression or classification tasks: recurrent neural network, convolution neural network, K-NearestNeighbor, and artificial neural network. As the performance of machine learning gradually increase, it’s interesting to study and examine its capability in medical field as well, specifically influenza trends prediction. In this study, this advanced technique would be implemented and tested using a real dataset from Center for Diseases Control and Prevention. Three different neural networks, recurrent neural network, long-short term memory, and gateway recurrent unit will be examined. This study would focus on comparing and contrasting the performance of each individual neural networks.

The rest of the paper is formatted as follow. In section 2, basic concepts would be briefly defined. In section 3, the structure of each model would be presented . In section 4, the effectiveness of each model would be analyzed and compared. In section 5, future work would be discussed.

**2 Basic concept**

*Definition 2.1* (Region). The whole United State is divided into 10 region when collecting data of outpatient illness surveillance:

Region 1—Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont Region 2—New Jersey, New York, Puerto Rico, and the U.S. Virgin Islands

Region 3—Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, and West Virginia

Region 4—Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee

Region 5—Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin

Region 6—Arkansas, Louisiana, New Mexico, Oklahoma, and Texas

Region 7—Iowa, Kansas, Missouri, and Nebraska

Region 8—Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming

Region 9—Arizona, California, Hawaii, and Nevada

Region 10—Alaska, Idaho, Oregon, and Washington

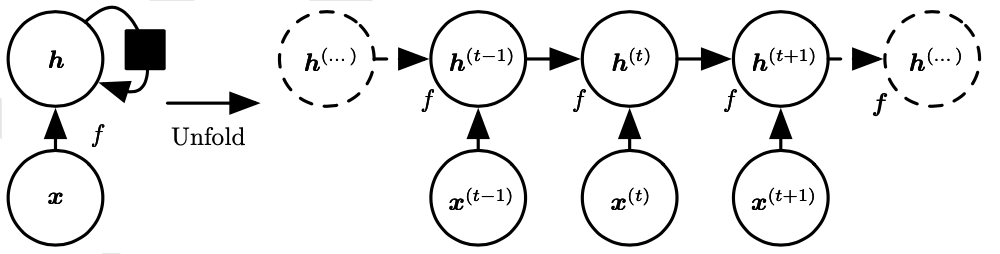
*Definition 2.2* (Time Series). Time series is a successive data point indexed by time sequence in which they occurred.

Here, Dwould be the entire time series collected from the CDC. It comprises all the previous weekly data point collected.

**3 Methodology**

**3.1 Recurrent neural network**

In this study, one of the deep learning model-Recurrent neural network (RNN)-is applied to make influenza prediction. A RNN layer is essential for RNN model as it selectively keep the relative aspects of the past sequence with better precision than the other. With the selected information, the model would then be able to predict the future trend.



**The graph of RNN Layer** (Goodfellow, Bengio and Courvil)

Here, x is the input value, t is the time step, and h is the hidden state. The h(x) hidden state is composed by the current input value x(x) and the previous hidden state h(x-1), which contains all the selected past information. Now, an equation can be used to represent the current hidden state after t time steps:

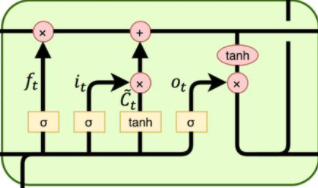
h(t)=f(h(t-1), x(t); )

Here, f would be the function that takes in all the past sequence(x(t), x(t−1), x(t−2),..., x(2),x(1)) as input to produce the current hidden state h.

Adding more RNN layer would of course make the model becomes more accurate. However, it’s too costly and unnecessary to have numerous RNN layers. One disadvantage of RNN is the fact that gradient for each time step may be lost when encountering long sequence. This would reduce the precision of the prediction of RNN

**3.2 Long-short term memory**

In this study, another deep learning model-Long-short term memory (LSTM)-is also implemented to do influenza prediction. A LSTM layer is pretty much similar to the RNN model, but even more complicated.



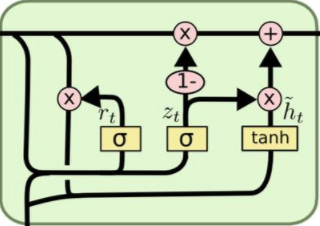
**The graph of LSTM layer** (jiang199912)

LSTM layer has additional three gateways to process data. Ft is the forget gate. The previous hidden state h(t-1) and current input x(t) would go through the sigmoid() activation function and obtains value between [0,1]. The closer the value is to 0, the greater chance the network would forget the previous input value. It is the input gate. The input data would go through both sigmoid and tanh activation to update the hidden state of the cell. Ot is the output gate. It outputs the h(t-1) and x(t) after sigmoid activation and contributes (Goodfellow, Bengio and Courvil) to the current hidden state h(t).

LSTM has the same function of capturing correlated data from the input sequence, but it’s better at storing gradient at each time step. LSTM layer is able to record all the gradient even with lengthy sequence of input. Gradient vanishing won’t occur as each gradient is used to do backward calculation to adjust every gateways. With this ability, LSTM performs generally better than RNN model in making future prediction.

**3.3 Gateway recurrent unit**

The last model implemented in this study is Gateway Recurrent Unit (GRU). It’s a variant of LSTM model, but simpler. This means that GRU has the same capability of LSTM: able to store the gradient at each step even with long sequence of input.



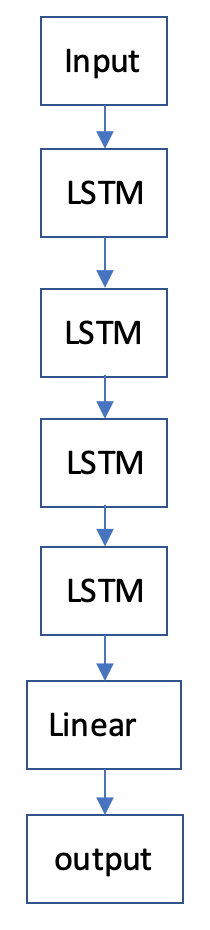
**The graph of GRU layer** (jiang199912)

GRU has only two gateways: update gate(zt) and reset gate(rt). Update gate is used to control the amount of data carried from the previous state to current state. Reset gate monitors the amount of data written into the backup file.

The advantage of GRU compare to LSTM is that it’s cheaper to operate and less time consuming. This is because there is less gateway and parameter to consider in GRU.

**3.4 Structure of the model**

The model is constructed with four LSTM, RNN, or GRU layers at the beginning. Then, it follows by a linear regression layer to produce the final prediction. In this study, I would like to use the data from the previous five months as input to predict the influenza percentage on the next month. So the input size for each LSTM, RNN, or GRU layer is set to be 5. The dataset would be reshaped into third dimension when inputting into LSTM, RNN, or GRU layer; then squeeze into two dimension before inputting into the linear regression layer. The linear regression would gather all the features produced by the previous four LSTM, RNN, or GRU layers and produce one output, which is the prediction of influenza for following month.

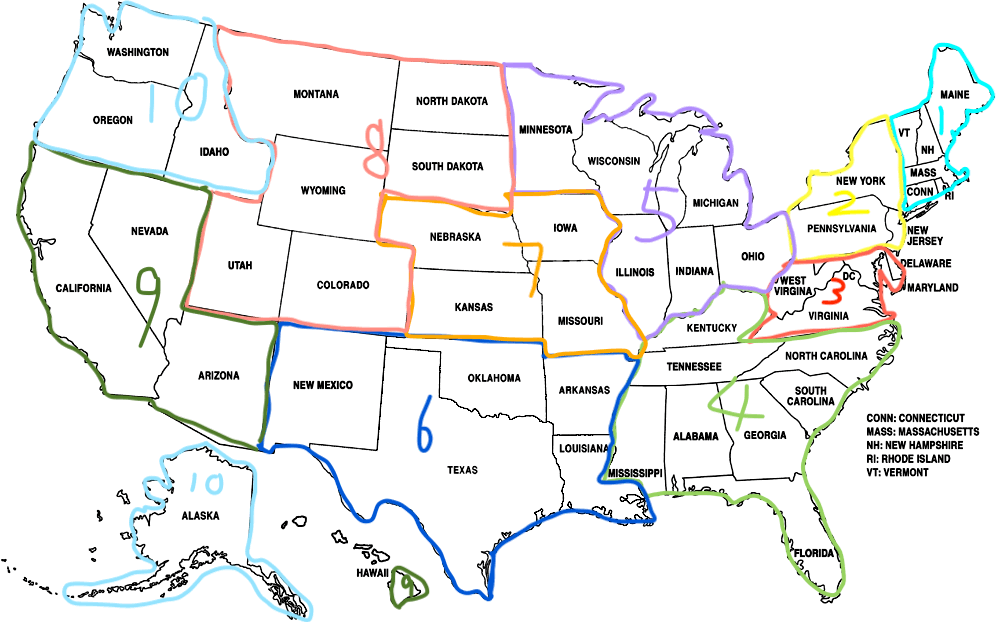


**4 Experiment**

Each model is evaluated using the real dataset from CDC. In this section, the used dataset would be introduced first. Then, a brief summary of data preprocessing would be presented. After that, the precision measurements would be exhibited. Finally, the result of each model would be presented and evaluated.

**4.1 Dataset**

The dataset used in this study is extracted from an digital reports from public health and clinical laboratories provided by the Center of Diseases Control and Prevention. The dataset includes the percent positive cases of outpatients in each week over the 10 regions starting from the 40th week of 2015 to the 46th week of 2018.

**The graph of 10 regions**

**Table 1: Dataset**

|  |  |
| --- | --- |
| Item | Comments |
| Location | 10 regions |
| Date span | 40th week of 2015 – 46th week of 2018. |
| Weeks | 163 |

**4.2 Data preprocessing**

Before inputting the dataset, all data points are set between the range of 0 and 1. The min-max normalization method is used to scale all the data points in the range. 80% of the data is used for model training and the rest of 20% data is used for validation.

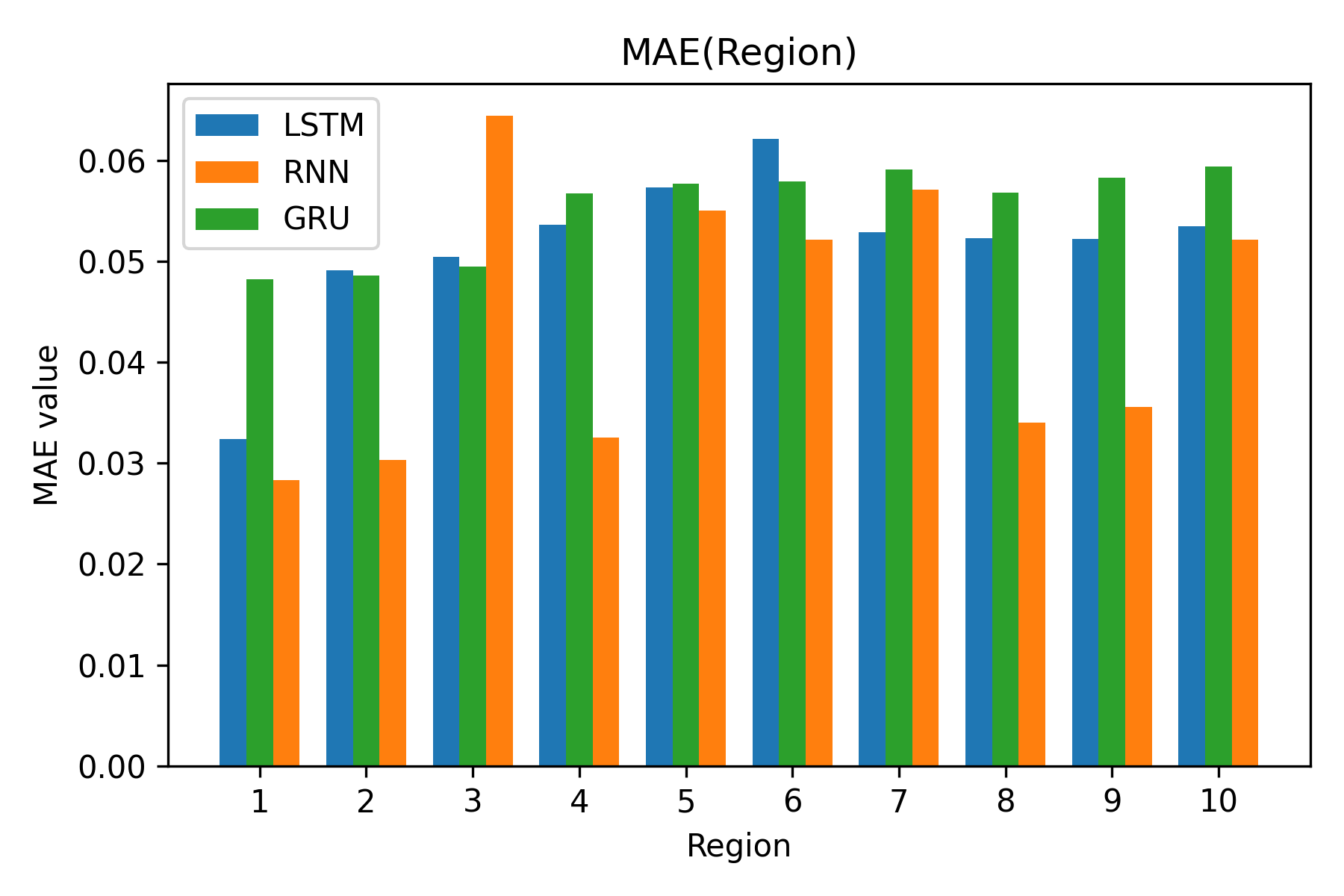
To better look at the performance of each model, both regional and national influenza datasets are used to train and predict.

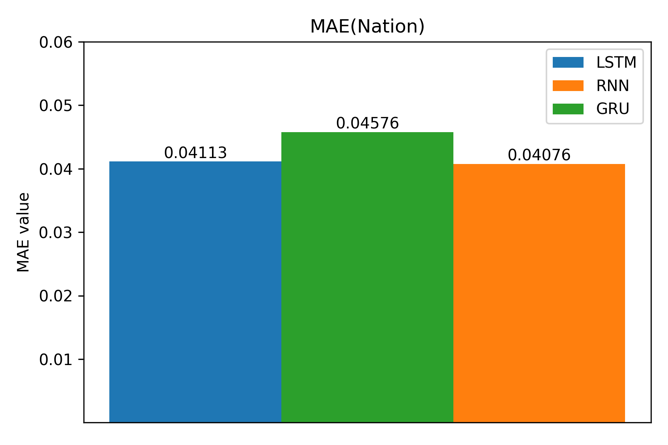
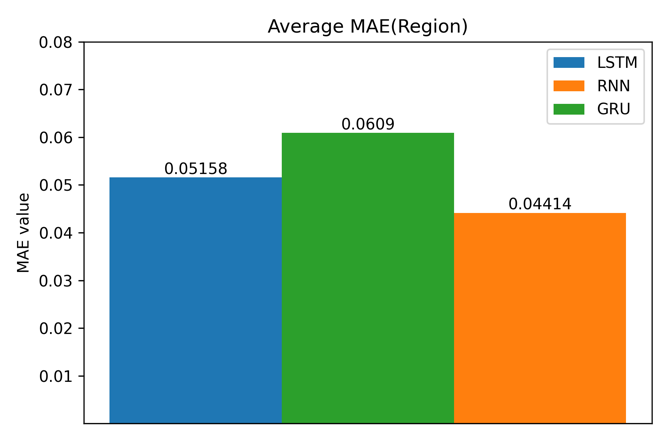
**4.3 Performance measurement**

To measure the effectiveness of each model, Mean absolute error(MAE) is calculated. The lower the value, the more accurate of the predicted values.

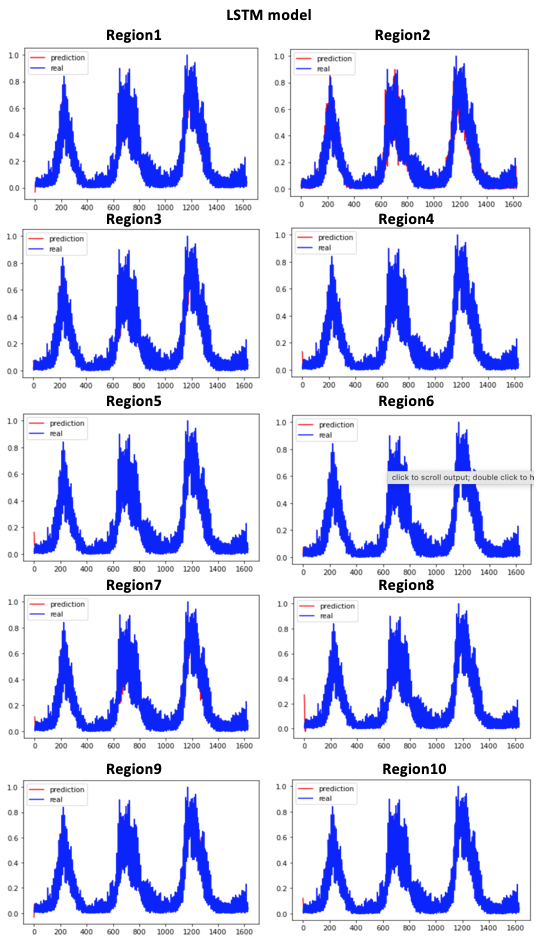
**4.4 Result evaluation**

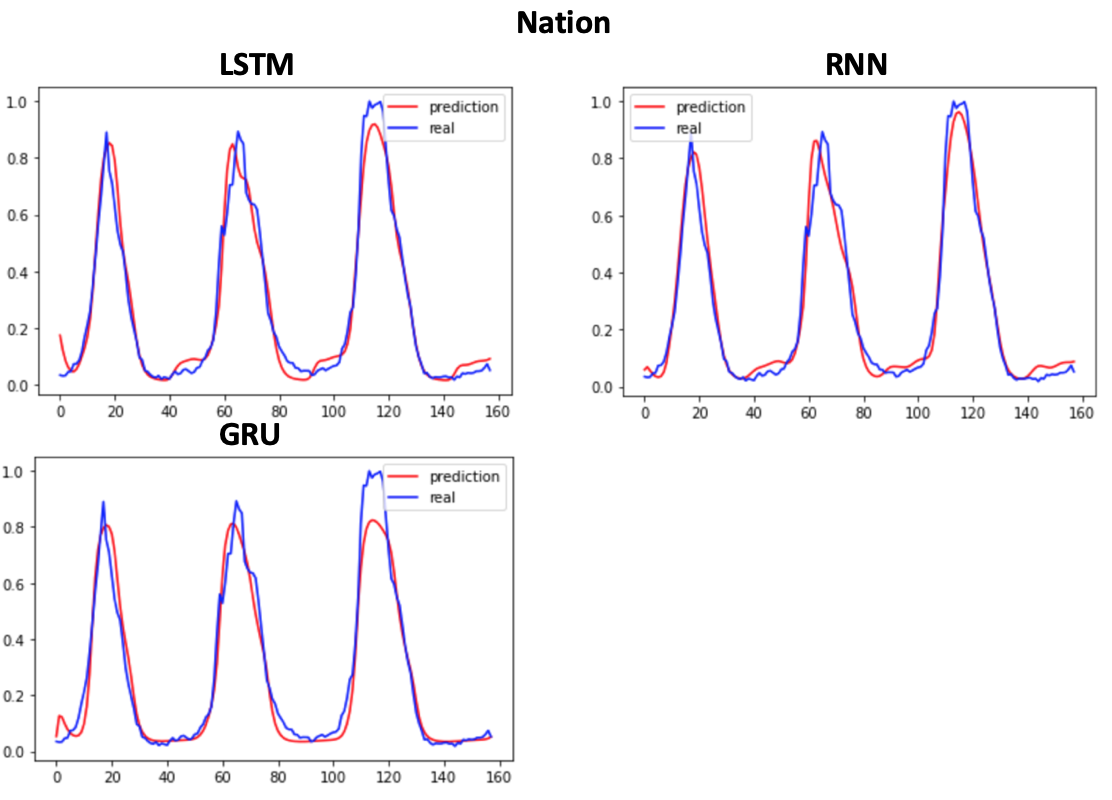
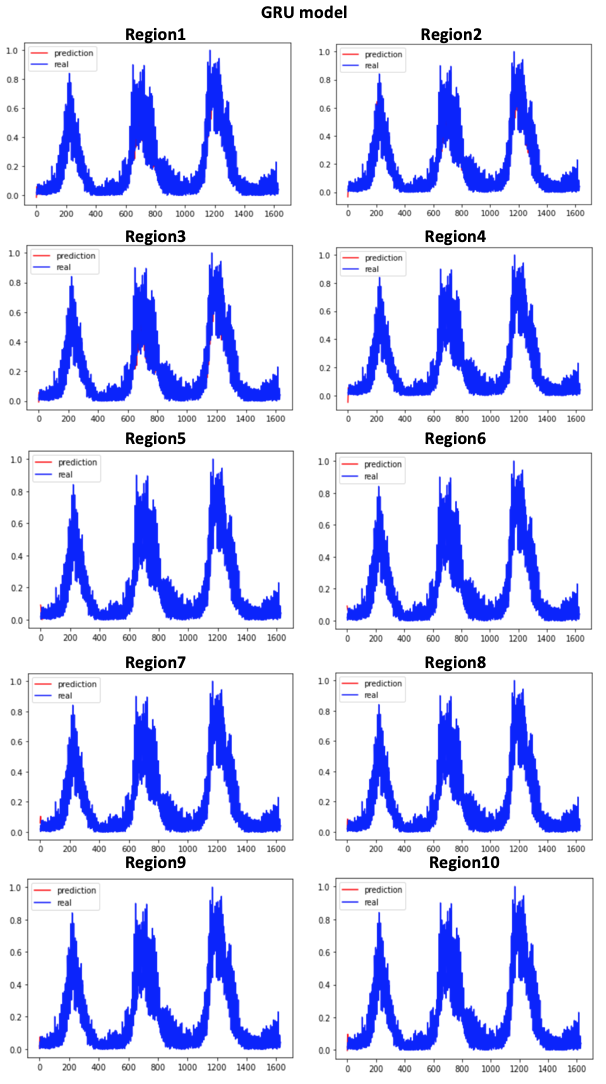
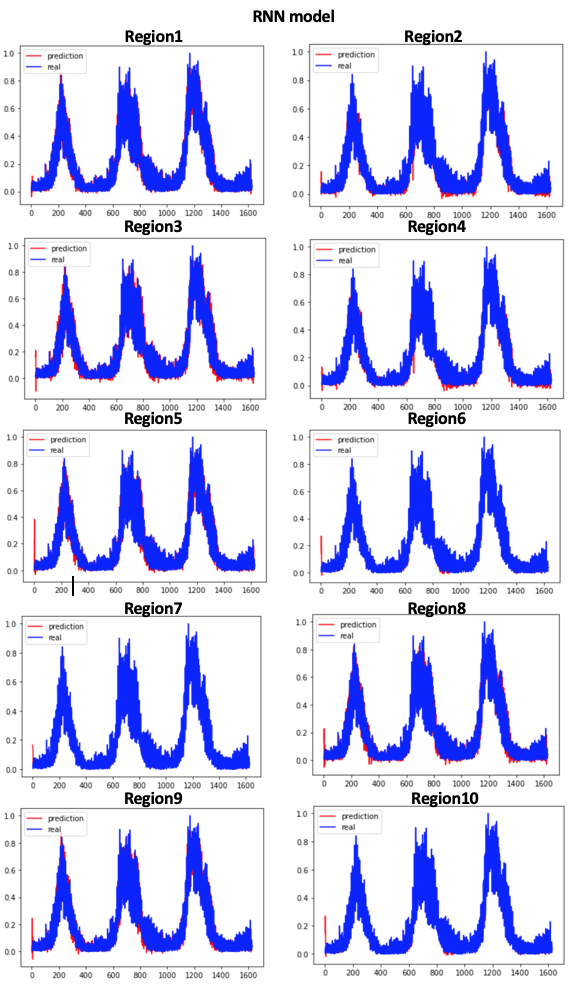
The data presented below come from the trained model after 100 epochs. Through comparison, RNN model is more than likely to have the best precision as it outperforms the other two baseline models in both regional and national influenza prediction. However, when the model is trained for 1000 epochs, LSTM model is more accurate in influenza trend prediction. This happens because the input sequence grows as more epochs is underwent by the model. A mentioned before, RNN starts to loss gradient as the input sequence becomes long. With missing gradient, the prediction made by the RNN model may contain greater deviations. In contrast, LSTM model wouldn’t encounter the problem of losing gradient after numerous time steps. With the entire gradient, LSTM model certainly would have better performance as the training time increases.





**Regional graph**

****

****

**5 Conclusion and future work**

The purpose of this study is to investigate the capability of various deep learning neural networks in predicting influenza trends. In this study, three different baseline models are tested on their ability to make accurate prediction on influenza trends. The experiment is conducted using real dataset from Center for Diseases Control and Prevention. The end results show that Recurrent Neural Network (RNN) has the best performance in making influenza trends prediction both regionally and nationally. Although LSTM and GRU models show less accuracy based on the MAE value, they are still competitive models in influenza trends prediction.

In the future, subtler and more advanced deep learning knowledge would be investigated and applied to the baseline model to improve the performances. Factors such as weather, temperature, location, and age group may be considered using Convolution Neural Network. Taking into account of this factors, the precision of the models’ predictions is more than likely to be enhanced.

**5.1 Appreciation**

Great thanks to both Dr. Kang Liu and Dr. Ruxin Wang of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. This study won’t be successfully accomplished without the programming and essay editing assistances and supervisions of the two professors.

# Bibliography

化龙十四. *循环神经网络*. 12 04 2021. 28 August 2021. <https://baike.baidu.com/item/%E5%BE%AA%E7%8E%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/23199490?fromtitle=RNN&fromid=5707183&fr=aladdin&ivk\_sa=1022817p>.

心雨叶子y and 沛 王. *长短期记忆人工神经网络*. 28 June 2021. 28 August 2021. <https://baike.baidu.com/item/%E9%95%BF%E7%9F%AD%E6%9C%9F%E8%AE%B0%E5%BF%86%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/17541107?fromtitle=LSTM&fromid=17541102&fr=aladdin&ivk\_sa=1022817p>.

微笑sun. *深度学习之GRU网络*. 27 July 2018. 28 August 2021. <https://www.cnblogs.com/jiangxinyang/p/9376021.html?ivk\_sa=1024320u>.

Goodfellow, Ian , Yoshua Bengio and Aaron Courvil. *Deep Learning*. 18 November 2016. August 2021. <https://www.deeplearningbook.org/>.

jiang199912. *lstm的理解*. 6 April 2020. 28 August 2021. <https://blog.csdn.net/jiang199912/article/details/105351831>.

Ltd, Elsevier . *The annual impact of seasonal influenza in the US: Measuring disease burden and costs*. 28 June 2007. August 2021. <https://www.sciencedirect.com/science/article/pii/S0264410X07003854>.

SAYA, MENGENAI. *United States Map Color Fill In*. 15 October 2017. 28 August 2021. <http://dafi1637.blogspot.com/2017/10/united-states-map-color-fill-in.html>.

Services, U.S. Department of Health & Human. *Types of Influenza Viruses*. 18 November 2019. August 2021. <https://www.cdc.gov/flu/about/viruses/types.htm>.

The Matplotlib development team. *Grouped bar chart with labels*. 13 August 2021. 28 August 2021. <https://matplotlib.org/stable/gallery/lines\_bars\_and\_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py>.