Neural-based epidemics prediction of COVID-19

Dexin Li 201905130191[1] and Yinchen Liu 201900130078[1]

[1]Shandong University, School of Computer Science

June 5, 2020

Abstract

The 2019-nCoV epidemic has totally changed our life. In order to limit its spread, every country has taken measures to limit public contacts and attempt to cure patients. Many of us are interested in how does the strict controls impact in the epidemic, and how soon will the society go back to normal. Instead of using heavy neural network structures or huge datasets, We propose a novel data pre-processing method, which help to get rid of some disadvantages of RNNs. We integrated epidemic data in China and deep-LSTMs to predict the data in worldwide. Our results show that the epidemic will probably converge by Aug.13, 2020.

1 Introduction

The corona-virus is one of the most contagious virus to hit our society. In December 2019, the outbreak of corona-virus occurred in Hubei Province, China. This epidemic explodes out rapidly. In January 21st, 2020, this virus has caused 262 infections. By May 5th, the confirmed data has risen to 3659271, surpassing the 2003 outbreak of SARS. In order to control the status, governments had carried out policies to limit physical contacts. Everybody wants to know how soon will the epidemic end. However, the policies are so complicated and the execution of people are so different, making the estimation difficult.

Time had witnessed the development of Artificial Intelligence(AI). Nanshan Zhong et.al [7] had modified the original SEIR model to emphasis the importance of dynamic Susceptible and Exposed population, which is state of the art. Including AI, Nanshan Zhong et.al also applied LSTM to the epidemic prediction. Deep Learning method had been used to predict various type of epidemic [3]. However, the principle of deep learning is hard to parse. Further more, deep learning takes a strict requirement of dataset, while epidemic data is full of noise and the quantity is not huge enough for deep network. Therefore, we used LSTM only for curve pattern exaction, just like Language Model(LM) in Natural Language Processing. We used a novel differential method to transfer curve prediction to LM. The training data comes from 2019 nCoV data in China, which turns out to be nearly convergent.

2 Data Smoothing

2.1 Data Source

We used confirmed, recovered and death data of COVID-19 in worldwide as validation data [4], and COVID-19 in China as training data. This is because the epidemic in China had almost come to an end. We also need data of SARS [6] for comparation.

2.2 Data Processing

This part is to describe what data structure we use for data interaction. After loading data, the function will return a two-dimentional-array. Each one-dimensional-array stores data for one country, whose element is an "Data" object represents the data of that country one day. So the whole Data structure is like this:

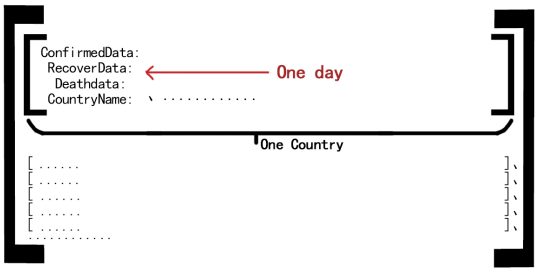
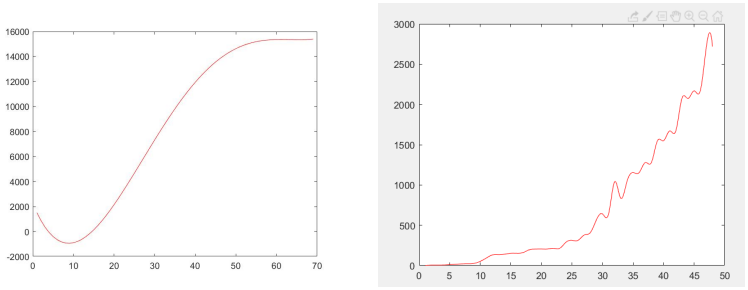


Figure 1: Data Structure

2.3 Curve Smoothing

The machine learning model requires high smoothness of the input data, and the curve formed by the original data does not meet the requirements, so curve smoothing is required. We tried three methods, polynomial fitting, cubic spline interpolation, and B-Spline.

After testing, we discoverd using the first two methords to process the data will produce large false oscillations.



(a) False oscillations of polyfit

(b) False oscillations of cubic spline interpolation

Figure 2: False oscillations

And finally decided to use the second-order B-Spline. Using this method, while ensuring that the result is relatively smooth, the result is highly compliant with the original data and basically does not produce false oscillations.

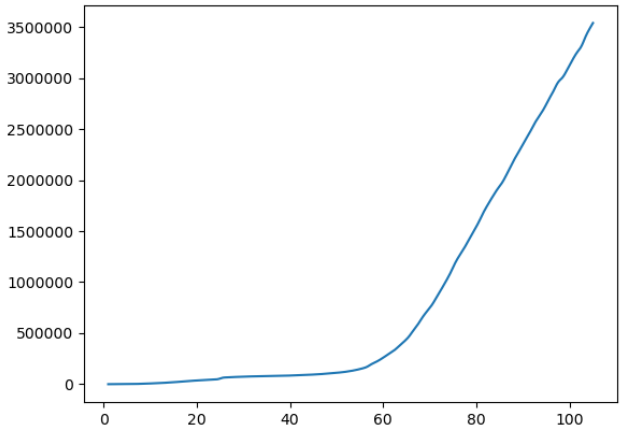


Figure 3: Worldwide data after smoothing

3 Preliminaries

The LSTM model is a type of RNN model which used forget gate and tanh layer to maintain its long-time memory. Deep LSTM turns out to be not interpretive. Generally, LSTM receives a series of input:

$$I = \{v_1, v_2, \cdots, v_n\}$$

, where v_i is a vector of a definite dimension m. An LSTM can be regarded a function defined in $\mathbb{R}^{m\times n} \to \mathbb{R}$.

Training is a process that fit the function into a dataset, so that the neural output of the input in dataset fits the dataset. The neural network will return an interpolation if the input does not come from the original dataset.

4 Mechanism

The original LSTM, which receives a huge raw dataset as training data(i.e. epidemic data of the past 20 years), turns out to be not such reasonable. The 2019 nCoV is more contagious than any epidemic before, and no interpolation between training set fits the status.

We notice that the infected curve goes to a definite shape. It is always an exponential curve when the virus or bacterial start to spread. As government notices the epidemic, the derivative of the curve starts to go down. In the end, the infected are controlled and the curve converges. It is the shape that determines the trend. Therefore, we only need to predict the curvature of the curve. However, the actual data is dispersed, which made the curvature impossible to compute. We used a substitute function for curvature. Define that:

$$C'(x,y) = \frac{\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}}{\frac{\mathrm{d}y}{\mathrm{d}x}}, x, y \in \mathbb{R}$$

Its dispersed form is:

$$\begin{aligned} C(x,y) &= \frac{\frac{(y_{x+1}-y_x)-(y_x-y_{x-1})}{2}}{\frac{y_{x+1}-y_{x-1}}{2}} \\ &= \frac{(y_{x+1}-2y_x+y_{x-1})}{y_{x+1}-y_{x-1}}, \\ x > 1, y \in \mathbb{R} \end{aligned}$$

This function uniquely determines a shape of curve. Formally, the shapes of two number sequence ϕ_x, θ_x are equivalent, when

$$\frac{\phi_x - \phi_{x-1}}{\theta_x - \theta_{x-1}} = \lambda$$

, where λ is a constant.

Theorem 1. *Given a series of rational $\{A_i | i \leq n, i \in \mathbb{N}, n > 1\}$, $C(x)$ determines a unique shape of number sequence.*

Proof. Sufficiency: Obviously, We can give the A_{n+1} by:

$$A_{n+1} = A_{n-1}C(n) - 4A_n + 2A_{n-1}$$

Necessity: Suppose there are two different ϕ_n and θ_n, if:

$$\frac{\phi_x - \phi_{x-1}}{\theta_x - \theta_{x-1}} = \lambda,$$
$$\frac{\phi_{x+1} - \phi_x}{\theta_{x+1} - \theta_x} = \lambda$$

, there is:

$$\begin{aligned} C(x,\phi) &= \frac{(\phi_{x+1} - \phi_x) - (\phi_x - \phi_{x-1})}{(\phi_{x+1} - \phi_x) + (\phi_x - \phi_{x-1})} \\ &= \frac{\lambda(\theta_{x+1} - \theta_x) - \lambda(\theta_x - \theta_{x-1})}{\lambda(\theta_{x+1} - \theta_x) + \lambda(\theta_x - \theta_{x-1})} \\ &= C(x,\theta) \end{aligned}$$

□

An LSTM [1] fetches out sequential features to predict a next data. Denote the LSTM units net^i receive a series of time-based input A_i, where A_i^{tr} stands for training data, and A_i^{in} stands for validation data. The expected output is denoted by net^{out}. A converged model should satisfy:

$$\exists A^{tr_1}, A^{tr_2} \in A^{tr}, s.t. \tag{1}$$
$$A_i^{in} \leqslant net^{out}(A^{tr_1}) \tag{2}$$
$$A_i^{in} \geqslant net^{out}(A^{tr_2}) \tag{3}$$

It turns out that the raw epidemic data does not mean to be available in legacy LSTM, while the preprocessed data fortunately satisfies the condition.

5 Rationality

A question is, why do we use the shape to determine a curve of epidemic? Does it fit the reality? In order to reveal its rationality, look at the two figures.

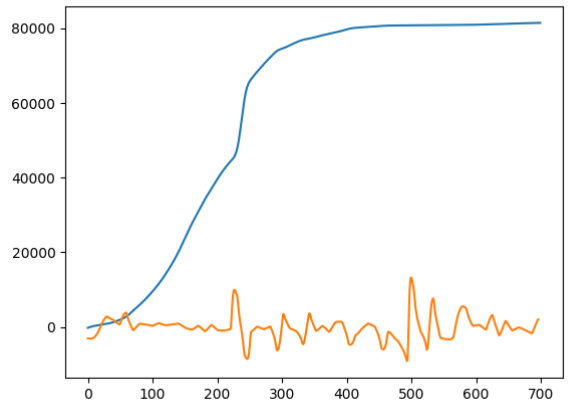


Figure 4: nCoV_China

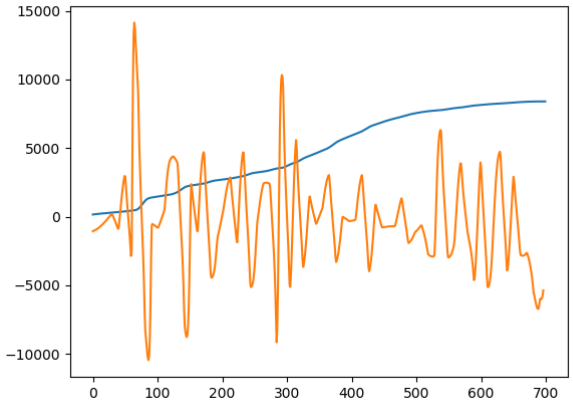


Figure 5: SARS

The orange curve shows $C(x, f)$, while the blue curve shows $f(x)$(notice that the curve has been smoothed and the x domain has been scaled to 10 times).

Notice that the $C(x, f)$ is basically an oscillation function. With the ending of epidemic comes, the vibration frequency and the amplitude are similar.

6 Prediction

Figure 5 shows the model prediction inputted with day $[50, 100)$, while Figure 6 shows another prediction inputted with day $[20, 70)$. The difference between the two figures reveals the different policies between worldwide and China.

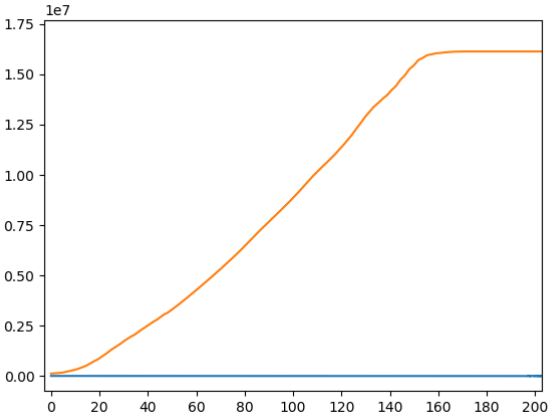


Figure 6: $[50, 100)$

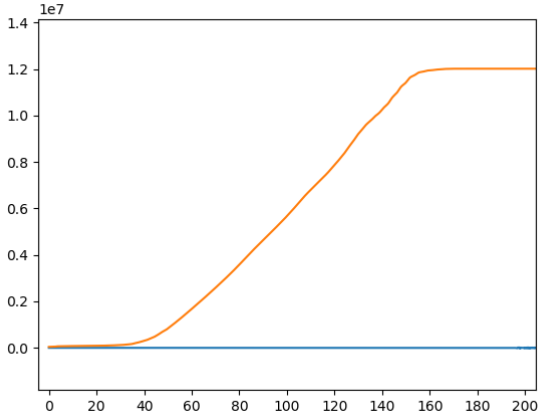


Figure 7: $[20, 70)$

7 Prospect

7.1 Epidemic curve under different policies

The epidemic has been raging for months. Here are the predictions for the United Kingdom and Brazil.

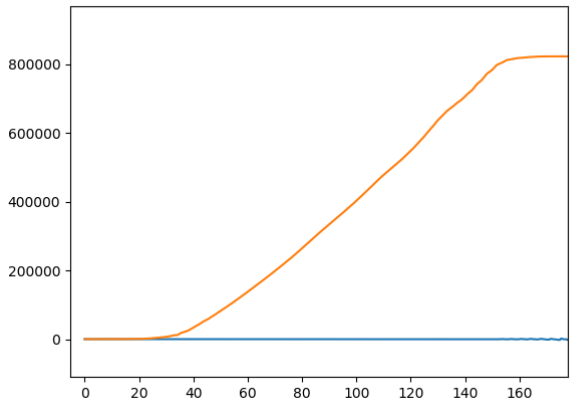


Figure 8: The United Kingdom

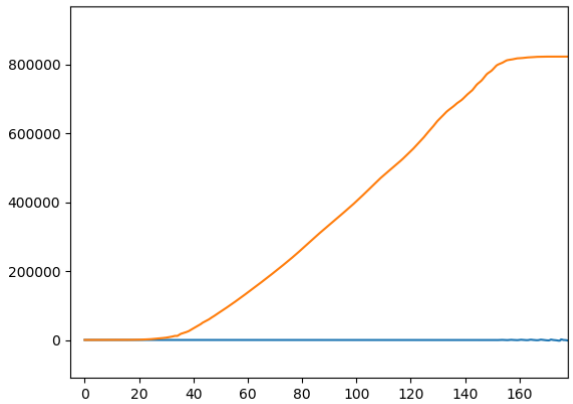


Figure 9: Brazil

The variants in epidemic predictions is countless. The most influential one is the policies. The United Kingdom turns out to rely on Herd immunity to

control the epidemic, which sacrifices too many infected. By contrast, China government has carried out effective policies, coordinated with people's execution, making the epidemic get efficacious control.

The classic SIR model [2], considered three compartments: susceptible, $S(t)$; infected, $I(t)$; removed, $R(t)$, where t stands for the time.

$$N = S(t) + I(t) + R(t) \tag{4}$$

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{5}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \frac{dR}{dt} = \gamma I \tag{6}$$

It is assumed that each person has equal possibility to be infected by a suspected each time. Thus, the infecting rate can be considered as a constant β, and radius as γ. As if we take measures to avoid being infected, such as wearing a mask, or avoiding people, both β and γ would be lowered. The epidemic would be controlled when it is in infancy. However, if β or γ can not be controlled, the exponential explosion will cause a disaster.

The explosion have been catching the governments' attention, and the infectious status have been cutting down.

7.2 Number of deaths in Brazil

We also made predictions on the number of deaths in Brazil. Accoding to the prediction above, we discoverd that the confirmed number of Brazil wiil peak around day 150, and the death toll of Brazil wiil reach about eighty thouthand. The curve is as follow.

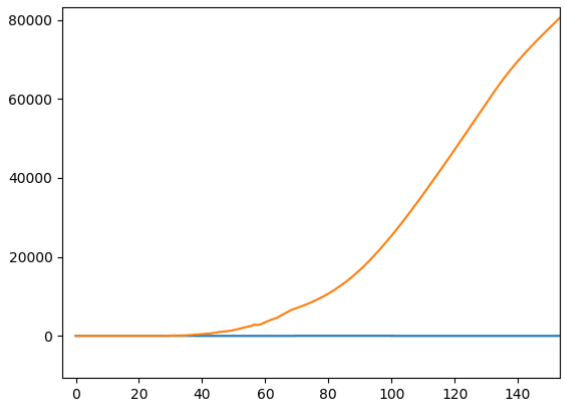


Figure 10: Number of deaths in Brazil

And the number of deaths worldwide will reach about a million by then.

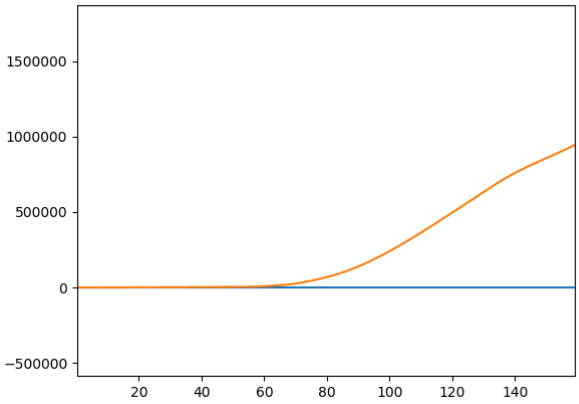


Figure 11: Number of deaths worldwide

7.3 Influence of medical resources

For the most severely affected cities, we did research on WuHan. Here is the confirmed of WuHan in February 2020.

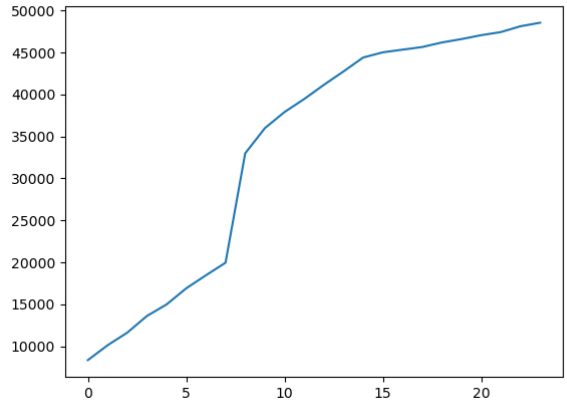


Figure 12: WuHan

From the figure we noticed that the confirmed number is increased very fast at the beginning, and slowed down at mid-February, when several mobile cabin hospitals were put into use.

The conclusion is very clear : the best way to slow down the spread of the virus is centralized isolation. Cities should establish a system to isolating infected people centrally when facing epidemic.If so, the pressure on the hospital will be greatly relieved, community and family infections can also be reduced.

7.4 Economic impact

The COVID-19 hit the economy hard. Accoding to the data provided by Nation Bureau of Statistics, the total tetail sales of consumer goods fell 15.8 percent and 7.5 percent year-on-year in March and April, respectively.

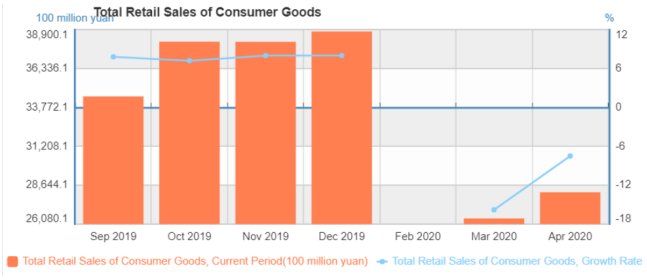
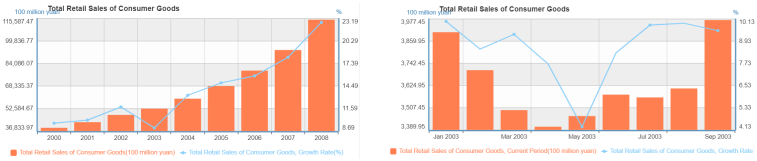


Figure 13: Total Retail Sales of Consumer Goods
[5]

Looking back on 2003, the situation is very similar. The outbreak of SARS slowed down the growth rate of retail sales of consumer goods in March, April and May 2003, and also greatly reduced the growth rate of it in 2003 the whole year.



(a) Influence in 2003 (b) Influence in 2003/5

Figure 14: SARS's influence
[5]

Although COVID-19 has a greater impact on the economy in the short term, observe the impact of SARS, we can find that the epidemic has little effect on the overall economic trend, so this time, we believe that the total retail sales of consumer goods this year may be lower than last year, but the overall stable situation will not change.

References

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

[2] William Ogilvy Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[3] Sijun Liu, Jiaping Chen, Jianming Wang, Zhuchao Wu, Weihua Wu, Zhiwei Xu, Wenbiao Hu, Fei Xu, Shilu Tong, and Hongbing Shen. Predicting the outbreak of hand, foot, and mouth disease in nanjing, china: a time-series model based on weather variability. *International Journal of Biometeorology*, 2017.

[4] Microsoft. Bing-covid-19-data. [EB/OL]. `https://github.com/microsoft/Bing-COVID-19-Data/tree/master/data` Accessed March, 8, 2020.

[5] National Bureau of Statistics. National data. [EB/OL]. `http://data.stats.gov.cn/english/ks.htm?cn=A01` `http://data.stats.gov.cn/english/ks.htm?cn=C01` Accessed June, 4, 2020.

[6] World Health Orgination. Cumulative number of reported probable cases of severe acute respiratory syndrome (sars). [EB/OL]. `https://www.who.int/csr/sars/country/en/` Accessed March, 11, 2020.

[7] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, Jingyi Liang, Xiaoqing Liu, Shiyue Li, Yimin Li, Feng Ye, Weijie Guan, Yifan Yang, Fei Li, Shengmei Luo, Yuqi Xie, Bin Liu, Zhoulang Wang, Shaobo Zhang, Yaonan Wang, Nanshan Zhong, and Jianxing He. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3), 2020.