

PROJECT STATUS REPORT

PROJECT NAME	Time Series Prediction for COVID-19	FINISHED BY	Yao Jiawei
--------------	-------------------------------------	-------------	------------

SUMMARY

This project is a simple time series prediction on the covid-19 dataset implemented by LSTM networks. It has been more than a year since the COVID-19 virus has ravaged the world, I think do a time series project by using COVID-19 dataset sounds like interesting. Hope that we can successfully defeat the virus.

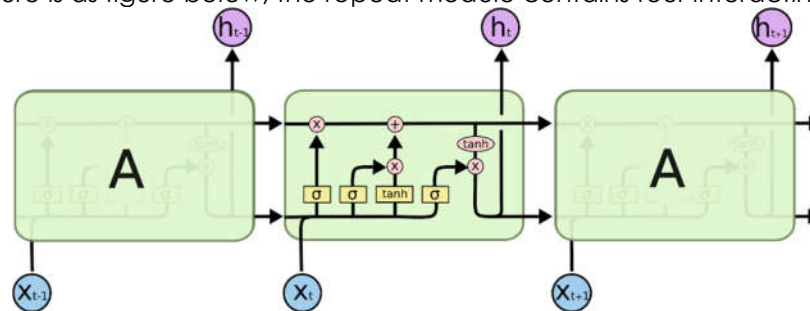
I choose the data from [COVID-19/2019-nCoV Time Series Infection Data Warehouse](https://github.com/BlankerL/DXY-COVID-19-Data) (<https://github.com/BlankerL/DXY-COVID-19-Data>) and I implement the project by LSTM(Long Short-Term Memory) networks.

The data I predict is the daily growth of american covid-19 cases.

INTRODUCTION OF LSTM

Long Short-Term Memory (LSTM) networks are a special type of recurrent neural network capable of learning order dependence in sequence prediction problems. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work.

A LSTM network structure is as figure below, the repeat module contains four interacting layers



The key to LSTMs is the cell state, the horizontal line running through the top of the diagram. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

Gates are a way to optionally let information through, thus it can restore the history information which may be related to current information.

The reason why I consider to implement this project by the LSTM is precisely the function of restoring the history information since I think the growth data will be affected by the history cases.

WORK PROGRESS

WEEK NO.	WORK	DETAILS
Week.1	Search for the dataset	Search data sets on covid-19 on github or other place on the internet, and I choose the COVID-19/2019-nCoV time series infection data warehouse, the data source is Ding Xiang Yuan which is a chinese medical community that collect the worldwid covid-19 data when the first covid case appears in Wuhan, China.
Week.2	Data process	Decide to filter out the american covid-19 data from the raw data to analysis. I selected the total count data whose country name is the United States from the original data set, and then calculate the difference between current day and the last day. This work is done by pandas.
Week.3	Implement an LSTM model	Implement an LSTM network model by Keras. There is only one layer of LSTM and then output to a dense layer to output result.
Week.4 & later	Solve some problems	Cause the result is very well for the prediction, I analyze the data and process. I think the problem is that the data fluctuates greatly. Thus I do the data smoothing by suing average value of seven days to indicate the value of the fourth day.

PROJECT RESULT

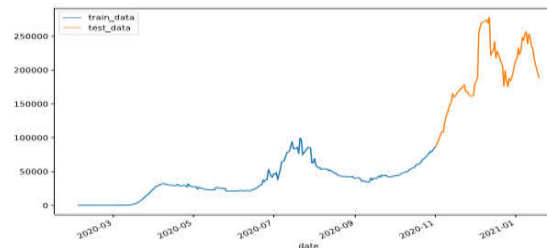


Figure 1 Training set and test set

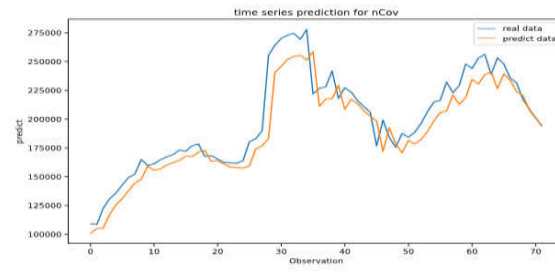


Figure 2 Prediction result and real data