LEARNING FROM THE PAST TO PREDICT THE FUTURE: APPLYING LSTM TO PREDICT COVID-19 PANDEMIC

Shenghui Chen
Department of Computer Science
University of Virginia
sc9by

Chang Cheng
Department of Computer Science
University of Virginia
cc8da

Sophia Cheung
Department of Computer Science
University of Virginia
xz3ts

July 16, 2020

ABSTRACT

Since January, the novel coronavirus outbreak in China has become a global pandemic, with more than 2 million and 789 thousand confirmed cases worldwide and over 200 countries and territories affected. As students in Machine Learning, we are interested in building a model to predict the spread of COVID-19 in Virginia, and some heavily affected countries and regions, such as Italy, Spain and New York, training, testing, and making predictions on the future growth trend of the number of confirmed cases in these areas. We use the models we built to figure out when newly affected areas will reach the peak, how many patients will be affected, and how long it will take for the outbreak to end.

1 Introduction

The new outbreak of Coronavirus has been the biggest news in the first month of 2020, especially for many of our friends who have families in the affected area. After WHO declared the situation in China a Public Health Emergency of International Concern, many countries have heightened their measures to better control the spread of the disease. This grave situation motivates us to study the spreading trend carefully, and contribute our share using knowledge we learned in the Machine Learning class. Specifically, we want to predict the trend of new confirmed cases by applying the LSTM model to the two major epidemic datasets (global confirmed cases and us confirmed cases).

We have done some literature search mainly in Machine Learning for epidemiological predictions, the machine learning methods for time-series data, and the existing mathematical models of infectious disease. In summary, we find many existing works on epidemiological predictions used RNNs as the main model [1]. One of the RNN model that works especially well with the time-series data is the LSTM model, which was applied in finance, e.g. to predict stock price [2]. Another existing model is a procedure for forecasting time series data called Prophet, which is open source software released by Facebook's Core Data Science team [3], though this model works best with time series that have strong seasonal effects. Finally, we get familiarized with the standard mathematical models of infectious disease such as SIR, SIRS, and SEIR.

2 Method

Based on the literature research on the prior work on forecasting time-series data, we have decided to use the Long short-term memory (LSTM) as our main model to train and test the data. Our data source is the Johns Hopkins CSSE dataset here: `https://github.com/CSSEGISandData/COVID-19`, where the collected data is from various sources including WHO, DXY.cn, China CDC, US CDC, etc.

For the overall approach, we intend to follow the steps listed below:

1. Data collection and cleaning to format it as inputs to a LSTM model
2. Discover and visualize data in different regions to gain insights

3. Preprocess the data through the same pipeline
4. Build a LSTM model and train it using data from region of interest
5. Fine-tune the model by comparing with actual data
6. Forecast the trend of spread in other newly affected regions

As of today, the three countries with the most confirmed cases of COVID-19 are the United States, Spain, and Italy. Inside the US, the region hit the hardest is the New York State. For this study, we aim to analyze the data from the most severely impacted regions and Virginia, which is most relevant to the UVa community, i.e. we will investigate the data in three countries: **the US, Spain, and Italy**, and two states inside the US: **New York and Virginia**.

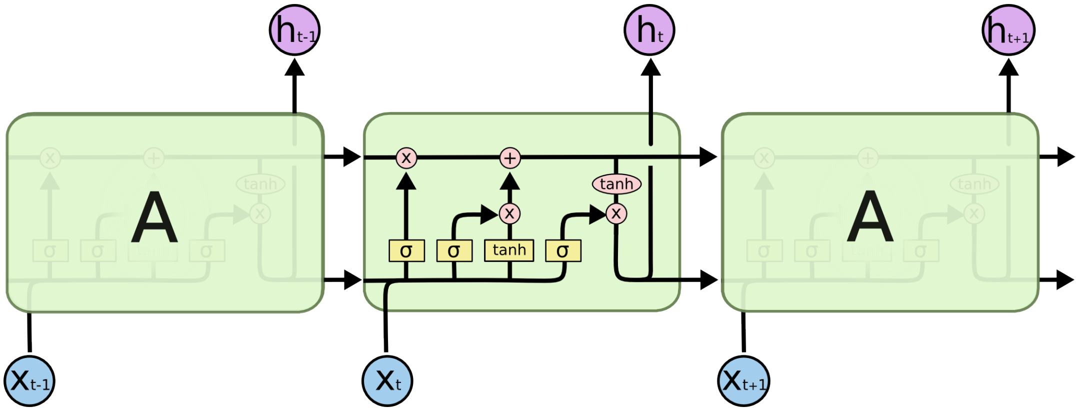


Figure 1: LSTM chain [4]

The Long Short-Term Memory (LSTM) model was proposed in 1997 by Sepp Hochreiter and Jürgen Schmidhuber and gradually improved over the years. It was designed to avoid the problem of long-term dependencies that cannot be handled by the standard RNN. The core idea behind LSTM is to have four interacting layers in each repeating module instead of a single one in that of a RNN module. These four layers learn to forget, store, update and output,

- Forget irrelevant parts of the previous state
- Store relevant parts in a long-term cell state
- Update selectively its cell state
- Output its cell state whenever needed

LSTM is well-suited to classify, process and predict time series given time lags of unknown duration. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, hidden Markov models and other sequence learning methods. In our study, we used the past data in respective countries and regions to train the LSTM model, and then use it to predict the future trends.

3 Experiments

We have conducted the experiment on **Google Colab**. We chose to use this platform because we can write and execute Python in our browser with zero configuration required, free access to powerful GPUs, and easy sharing with each other.

For this project, we chose the **Johns Hopkins CSSE** data source because it is the most comprehensive, open-source platform of confirmed COVID-19 cases around the world, where the collected data is from various sources including WHO, DXY.cn, China CDC, US CDC, etc.

For LSTM models, we have tried several models with variation in number of layers and neurons, and different activation functions and optimizers. The LSTM models that yield the best predictions results have three LSTM layers accompanied with two dropout layers that have dropout rate between 0.2 to 0.6, and use 'relu' and 'elu' activation functions with 'adam' optimizer. When we forecast the future spreading of coronavirus, we have tried to use the number of confirmed cases in the past 10 days to 1 day to predict the confirmed cases in the next 1 to 5 days. The results show that using the number of confirmed cases in the past 4 days to predict the number of confirmed cases in the next two days yields the best predictions.

Based on the results of the experiments, we can derive the following outcomes:

- Spain, Italy and the New York state is predicted to reach a plateau within one month.
- The United States as a whole and the state of Virginia is still in the phase of rapid increasing confirmed cases, and is predicted not to reach a peak within one month.

To assess the robustness of the algorithm, we have plotted the loss function and accuracy for 100 epochs for different regions. Since all five models have similar accuracy and loss functions, we take the plot for the United States data as an example (Figure 2).

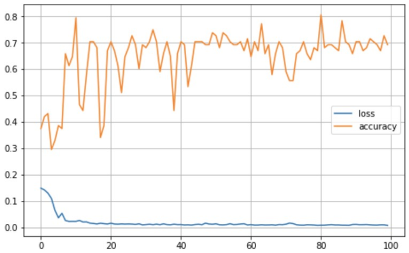
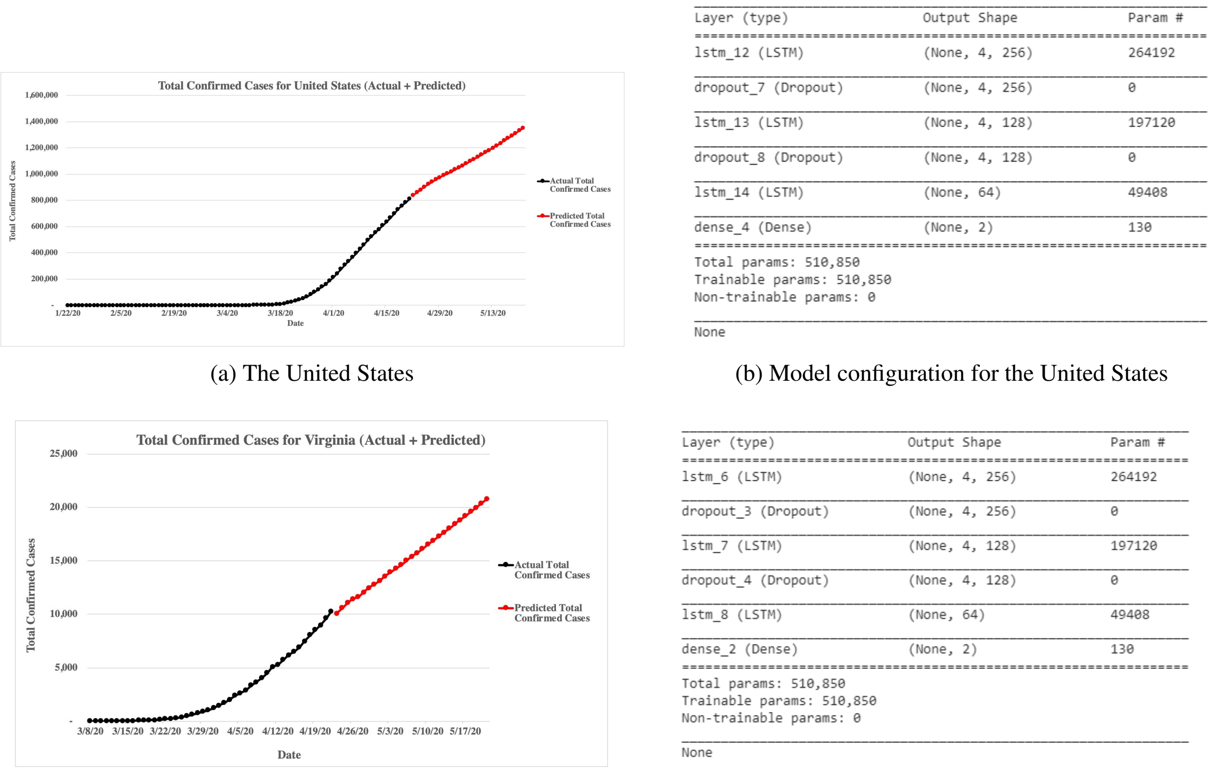


Figure 2: Loss function and accuracy of the United States for 100 epochs

4 Results

Based on the experiment we conducted above, we have results for each region shown in the figures below. From the figures we can see the trend for each region and plan for the following stage of the outbreak based on the prediction given. All relevant files and figures of this project is contained in this github repository: `https://github.com/viviangchen98/coronavirus_prediction`



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_12 (LSTM)               (None, 4, 256)            264192
_________________________________________________________________
dropout_7 (Dropout)          (None, 4, 256)            0
_________________________________________________________________
lstm_13 (LSTM)               (None, 4, 128)            197120
_________________________________________________________________
dropout_8 (Dropout)          (None, 4, 128)            0
_________________________________________________________________
lstm_14 (LSTM)               (None, 64)                49408
_________________________________________________________________
dense_4 (Dense)              (None, 2)                 130
=================================================================
Total params: 510,850
Trainable params: 510,850
Non-trainable params: 0
_________________________________________________________________
None
```

(a) The United States

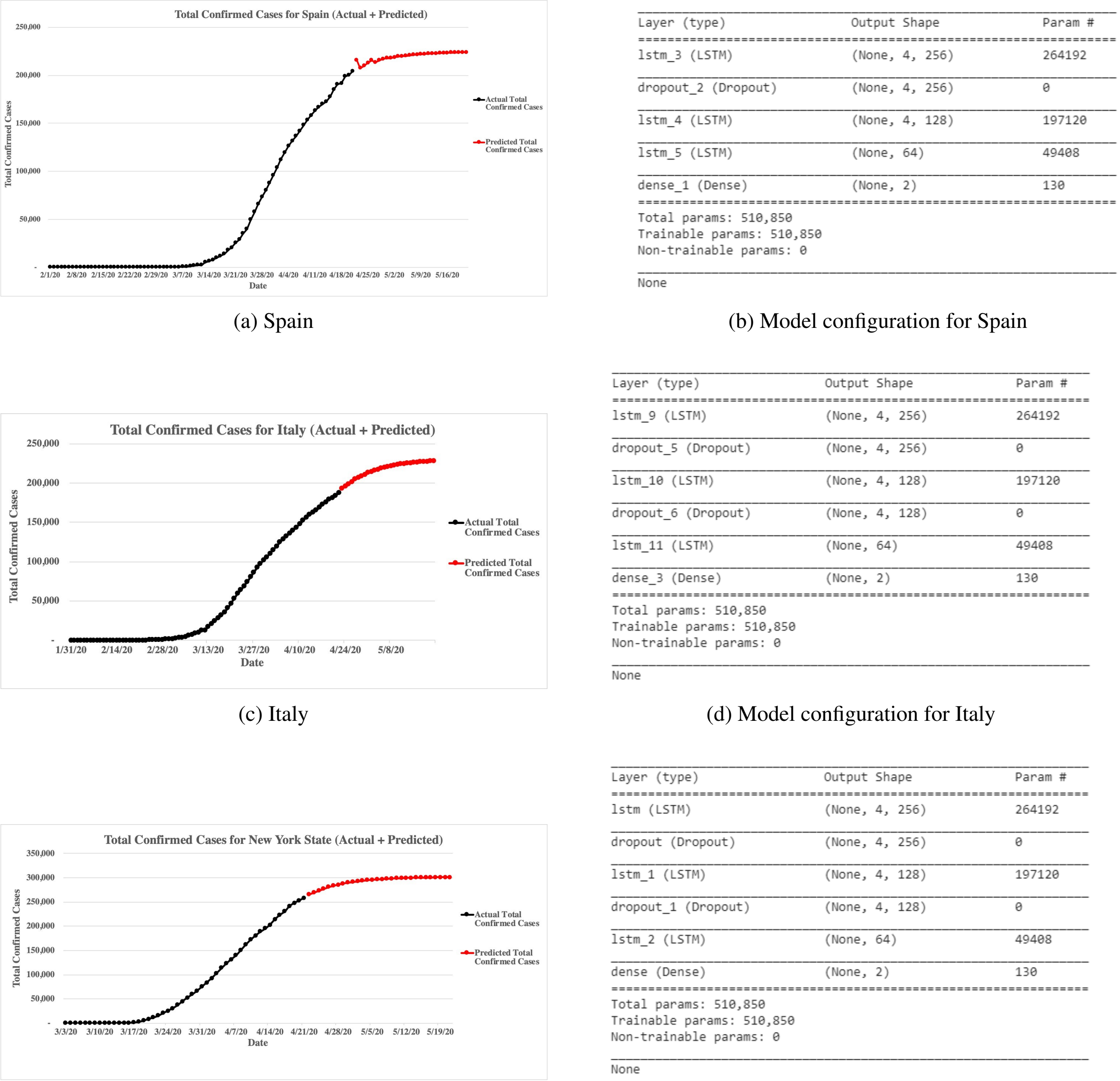
(b) Model configuration for the United States

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_6 (LSTM)                (None, 4, 256)            264192
_________________________________________________________________
dropout_3 (Dropout)          (None, 4, 256)            0
_________________________________________________________________
lstm_7 (LSTM)                (None, 4, 128)            197120
_________________________________________________________________
dropout_4 (Dropout)          (None, 4, 128)            0
_________________________________________________________________
lstm_8 (LSTM)                (None, 64)                49408
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 130
=================================================================
Total params: 510,850
Trainable params: 510,850
Non-trainable params: 0
_________________________________________________________________
None
```

(c) Virginia

(d) Model configuration for Virginia

Figure 3: The total confirmed cases with prediction for regions still in rapid increase



(a) Spain

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 4, 256)            264192
_________________________________________________________________
dropout_2 (Dropout)          (None, 4, 256)            0
_________________________________________________________________
lstm_4 (LSTM)                (None, 4, 128)            197120
_________________________________________________________________
lstm_5 (LSTM)                (None, 64)                49408
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 130
=================================================================
Total params: 510,850
Trainable params: 510,850
Non-trainable params: 0
_________________________________________________________________
None
```

(b) Model configuration for Spain

(c) Italy

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_9 (LSTM)                (None, 4, 256)            264192
_________________________________________________________________
dropout_5 (Dropout)          (None, 4, 256)            0
_________________________________________________________________
lstm_10 (LSTM)               (None, 4, 128)            197120
_________________________________________________________________
dropout_6 (Dropout)          (None, 4, 128)            0
_________________________________________________________________
lstm_11 (LSTM)               (None, 64)                49408
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 130
=================================================================
Total params: 510,850
Trainable params: 510,850
Non-trainable params: 0
_________________________________________________________________
None
```

(d) Model configuration for Italy

(e) New York state

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 4, 256)            264192
_________________________________________________________________
dropout (Dropout)            (None, 4, 256)            0
_________________________________________________________________
lstm_1 (LSTM)                (None, 4, 128)            197120
_________________________________________________________________
dropout_1 (Dropout)          (None, 4, 128)            0
_________________________________________________________________
lstm_2 (LSTM)                (None, 64)                49408
_________________________________________________________________
dense (Dense)                (None, 2)                 130
=================================================================
Total params: 510,850
Trainable params: 510,850
Non-trainable params: 0
_________________________________________________________________
None
```

(f) Model configuration for New York state

Figure 4: The total confirmed cases with prediction for regions about to reach a plateau

5 Conclusion

In this project, we attempt to predict the trend of number of confirmed cases of COVID-19 in severely affected regions by training a Long Short-Term Memory model from historical data. We have implemented the data extraction, model creation, training and predictions in Google Colab. The results show that Spain, Italy and New York state is reaching toward a plateau albeit at varying scale, while the United States as a whole and the state of Virginia in particular is still expecting more increase in confirmed cases for a period of time. This has significant social and economic impacts on the well-being of the commonwealth of Virginia, as well as the entirety of the United States, as the continuous spread of the virus may not only lead to shortage in medical resources, but may also permanently damage the economy as more small businesses shut down. In light of these discoveries, we can potentially plan ahead and allocate scarce resources across different regions based on the real-time need.

This is a very preliminary study, and one of the limitation is the lack of consideration for a change in behavioral patterns in a region. We recognize that a model is as good as its assumptions. The model we built learns from the historical data in a region, and the assumption we made to predict future trend based on this data is that people maintain the same behavioral patterns (e.g. no stricter regulations, no lockdown of a city). However, this assumption is hard to forecast. Therefore, the more careful epistemological study of the COVID-19 spread surely needs to be a dynamic analysis that involves multiple social groups including scientists, governments, and the citizens. Given the severity of the issue, any future study will need to ground itself in the most urgent need in the society.

6 Contribution from each member

In this project, all group members have contributed equally to the project, with the rough division of tasks detailed below:

- Shenghui Chen: literature research, finding data sources, writing documentations
- Sophia Cheung: data collection and pre-processing, building up the LSTM model
- Chang Cheng: data visualization, training and testing of the LSTM model

References

[1] Yuexin Wu, Yiming Yang, Hiroshi Nishiura, and Masaya Saitoh. Deep learning for epidemiological predictions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1085–1088, 2018.

[2] Kai Chen, Yi Zhou, and Fangyan Dai. A lstm-based method for stock returns prediction: A case study of china stock market. In *2015 IEEE international conference on big data (big data)*, pages 2823–2824. IEEE, 2015.

[3] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[4] Christopher Olah. Understanding LSTM Networks.