Time series analysis of COVID-19 cases using deep learning models : LSTM & CNN

Sakshi Joshi

Master of Science in Business Analytics, Michigan State University, East Lansing 48823, US

***Abstract* – COVID-19 has been a major threat to every aspect of human life. It has deeply affected the global market as well as personal lives. In this paper, Deep Learning-based models are used for predicting the number confirmed and death cases due to Covid-19 globally. The model can predict COVID-19 cases for 159 countries. Recurrent neural network (RNN) based long-short term memory (LSTM) and Convolutional Neural Network are used to build a model on the global COVID-19 dataset to predict the number of confirmed cases and fatalities. By experimenting will combinations of LSTM and Convolutional LSTM, it was clear that a combination of stacked LSTM and Conv layers gives better predictions compared to LSTM model alone. The accuracy metric used was mean square error (MSE) and hyperparameter tuning was considered to build a robust model.**

***Keywords* – LSTM, CNN, Time Series, COVID-19**

I. INTRODUCTION

The COVID-19 outbreak first occurred in the city of Wuhan, China in December 2019. The cause of this outbreak is not 100% proven although there are a lot of theories claiming Corona virus (the family of virus responsible for COVID-19, SARS, etc.) originally came from Bats and snakes. This highly transmitted virus gradually spread around the world. The major medium of the spread of this virus is been anticipated as human contact or bodily fluid transfer. The governments all across the world have imposed restrictions on human contact. The year 2020 has seen the unthinkable with more than 80% of the countries under lockdown, new work-from-culture, heavy layoff, ban on travelling, stay-at-home order, etc. The world was not ready for such a pandemic and if not for vaccine, many country's economy will be crashing soon. Researchers across the world are trying to use analytics and machine learning to understand the indicators driving this COVID-19 outbreak. There are a few analyses available online and few great repositories tracking daily COVID-19 outbreak globally. One such platform is the John Hopkins Coronavirus research center [1]. This website tracks daily confirmed cases, death cases and recovered cases across the globe. It has data of 189 countries, out of which 8 countries even has statewise and countywise data. This data is available from January 22nd , 2020 till today.

To cater to this time series data, Time series analysis using LSTM models can be used. LSTM, Long Short-Term Memory, is a type of Recurrent Neural Network and has gained great attention due to its vast time series applications like weather prediction, stock price prediction. LSTM was proposed to overcome the weakness of RNN - dealing with long-sequence data insisting the inability to handle the vanishing gradient problem during the learning process [2]. LSTM, which contains the input, output, and forget gate to better capture the correlation of data with long term dependencies [3]. The LSTM parameter, however, needs to be

optimized depending on data characteristics by choosing the number of the layers or hidden units, especially for highly complex data, which are non-linear and long. In this paper, we propose an LSTM framework that handle the nonlinearity and complexity of COVID-19 time-series data. This paper proposes enhancement of an existing framework which uses LSTM layers in the model. The final model consists of LSTM and CNN framework with a layer of Convolutional LSTM with activation function ReLu, followed by layers of LSTM cells (stacked LSTM), followed dense layers and dropout for regularization. The layers are then tuned with respect to the training time series data. For performance boosting, we can keep adding or changing hidden layers to find the best results. Each LSTM layer has different resolutions to get the best performance.

One important practice in the paper is that the model is built using external data. This external data is taken from official websites which will help the model to understand the key indicators in predicting the patterns in COVID-19 cases. This external data, also time series, is fed to the Convolution LSTM model sequential using multi input LSTM model. The input layers are then merged into one layer and the LSTM layers are followed by this concatenated layer. Details regarding this model and the dataset used will be explained in the following sections.

Section II talks about the basic Convolutional LSTM model which takes time series COVID-19 data as input. This analysis is done on the cumulative global COVID-19 data. Section III introduces the unique features of the enhanced model and Section IV explains the multi input Convolutional LSTM model along with all its characteristics and datasets used. Section V talks about the results of both the models. Section VI talks about the conclusion and future scope of the study.

II. BASELINE MODEL

The baseline model is referred from the github jupyter notebook[4] that predicts cumulative COVID-19 active and recovered cases, and death count. The dataset used here is the JHU COVID-19 time series data which has daily case counts for 189 countries.

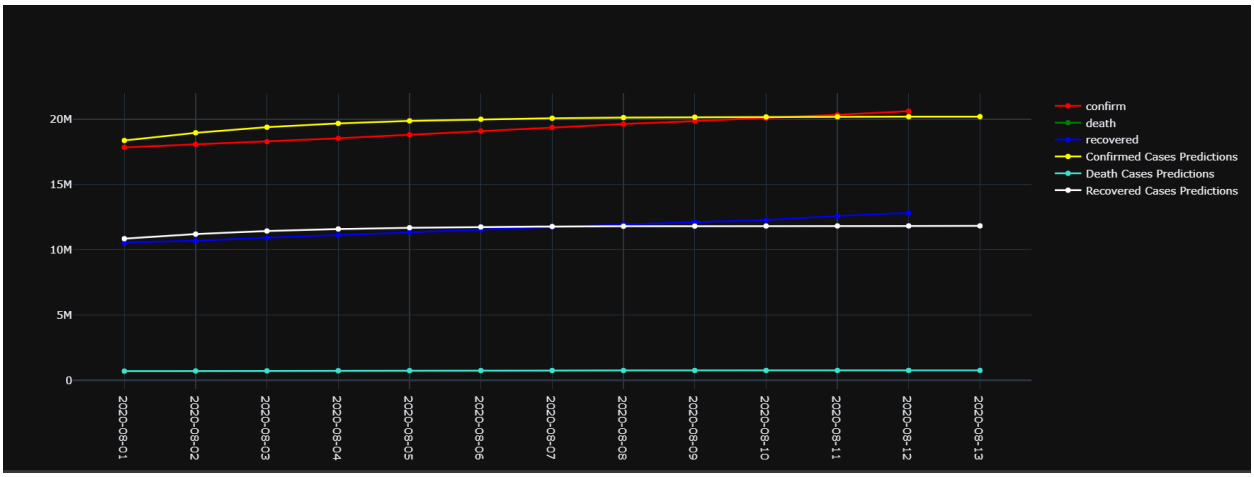
The dataset is in time series format with every day COVID-19 data on a country level. This dataset acts as both training data and testing data to evaluate the performance of the model. The baseline model an input layer which takes time series data from January 22nd , 2020 till today (the JHU data repository is updated every day for new COVID-19 data points). This layer is am LSTM layer. The LSTM is followed by a dropout layer with dropout rate 0.2 , followed by the output layer, a dense layer with 1 filter. This model is compiled using Adam optimizer and loss function as mean squared error (MSE). The model is fitted for the given training and testing dataset with batch size 16 and number of epochs as 500. The baseline model architecture is shown below.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 1, 128)            66560
_________________________________________________________________
dropout (Dropout)            (None, 1, 128)            0
_________________________________________________________________
dense (Dense)                (None, 1, 1)              129
=================================================================
Total params: 66,689
Trainable params: 66,689
Non-trainable params: 0
_________________________________________________________________
```

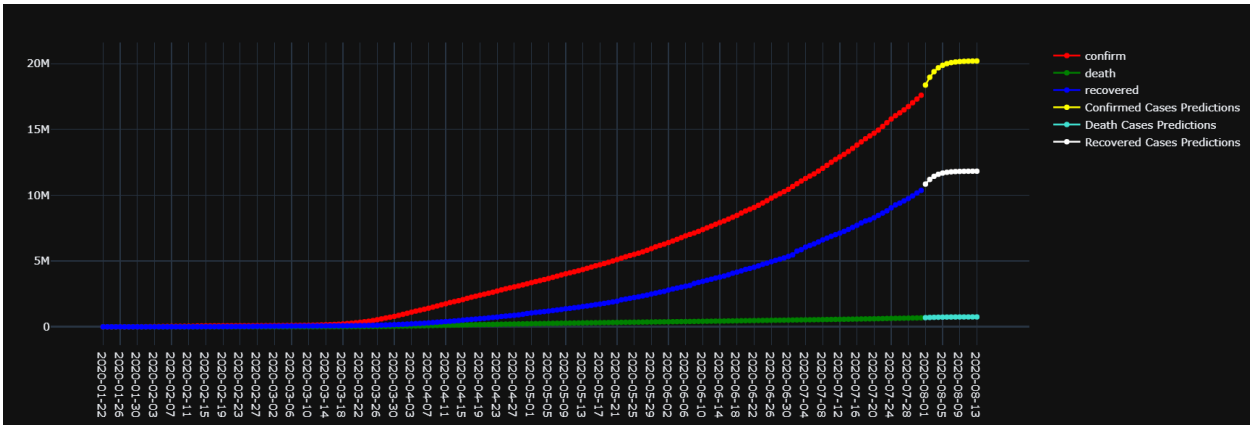
Baseline model

The above model is trained for all three datasets individually – confirmed cases, death count and recovered cases.

Then, depending on the user given number of days to predict future cases, the model is run in loop for those many times. The output of all three models are shown below.



Baseline mdoel Prediction for confirmed, death and recovered cases



Visualization with historical COVID-19 data

This is the output of the model when the model is trained till 31st July, 2020 data and predicted cases for 1st August, 2020 till today. If you look closely, the prediction plateaus after few initial iterations. This is the major flaw of the model. This is because the time step of the model is 1, i.e. the model considers just 1 day prior day to make future prediction. This is highly unreliable since COVID-19 cases depends on historical data as well as other parameters.

The LSTM model is featured to consider historical data and this functionality can be used by increasing the time steps of the model. Furthermore, this model is unreliable because it does not consider external data like country deomgraphics, government responses, popluation, etc. These parameters highly influence the outbreak pattern of COVID-19.

In order to implement this, we have engineered an enhanced model. This model will show better results compared to the baseline model.

III. ENHANCED MODEL FEATURES

The model engineered in this paper has a lot of new characteristics compared to the baseline model. Each of these characteristcs are explained in the following sub sections.

A. External Dataset

As mentioned earlier, COVID-19 outbreak is influenced many parameters including the population of the country, medical facilites, citizen demographics, government responses, etc. To cater to a few of these features, the model using external time series data from different official sources.

Government Response –

The most import aspect of COVID-19 pandemic is human contact and to decrease this, goverments around thr world have taken serious actions to control the outbreak. There is a time series data tracking government responses of 161 countries[5]. This is an extensive dataset with 17 indicators including containment & closure policies, economic policies and health system policies. Out of these 17 indicators, the model uses a mixture of 9 indicators as follows-

School_closing
workplace_closing
cancel_public_events
restrictions_on_gatherings
close_public_transport
stay_at_home_order
restrictions_on_internal_movement
international_travel_control
public_information_campaign

These 9 indicators are 9 separate time series datasets and using this will be very expenisve computional wise. Therefore there is an index, called Stringency Index, which calcutes an index for every country. The formula for this is below,

$$(1) \qquad index = \frac{1}{k}\sum_{j=1}^{k} I_j$$

Here, K is 9 since there are 9 inidicators. This new Stringency dataset is a single time series dataset which can now be used in any LSTM model.

Population Density –

Another important indicator that indicates the risk of COVID-19 outbreak is the population of any country. With more population comes greater danger of an outbreak. More importantly a country's population density will be more useful than the population as density gives a better idea of a country's human capital. This density data is extracted from [6], a 2019 prediction of Censes 2010 data, which gives population density for 159 countries. This is however is not a time series data. This dataset is given seperatly to the LSTM model.

This way, the model now has three different data input – COVID-19 data, Government Response data and Population Density data.

B. Scaling Dataset

Another important thing to be considered in the new datasets is the variation amongst the data points. Countries like US, India, Brazil which have high COVID-19 numbers can skew the model. This needs to be addressed by adding a scalling method in the data processing stage. For this model, MinMaxScaler and RobustScaler are used. These methods scales down and normalizes the data to nullify the effect of outliers.

C. Convolutional LSTM Model

The LSTM time series model gives much better results when a Convolutional layer is added to the model. This layer is added after merging the 3 input layers explained above.

IV. ENHANCED MODEL

The data sets given to this model are explained in the above section. Once these datasets are fetched and processed, we build the architechture of the model. Refer the below image to understand the architecture of the moel.

Model: "functional_1"

Layer (type)	Output Shape	Param #	Connected to
cases (InputLayer)	[(None, None, 1)]	0	
external (InputLayer)	[(None, None, 1)]	0	
pop (InputLayer)	[(None, None, 1)]	0	
concatenate (Concatenate)	(None, None, 1)	0	cases[0][0] external[0][0] pop[0][0]
conv1d (Conv1D)	(None, None, 16)	48	concatenate[0][0]
lstm (LSTM)	(None, None, 30)	5640	conv1d[0][0]
dense (Dense)	(None, None, 15)	465	lstm[0][0]
dropout (Dropout)	(None, None, 15)	0	dense[0][0]
lstm_1 (LSTM)	(None, None, 30)	5520	dropout[0][0]
dense_1 (Dense)	(None, None, 1)	31	lstm_1[0][0]

Total params: 11,704
Trainable params: 11,704
Non-trainable params: 0

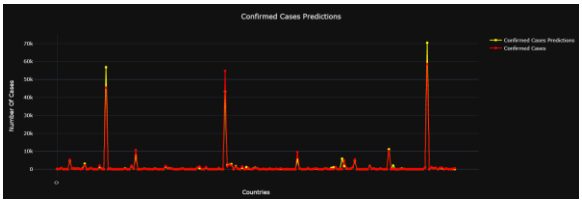
Model Architecture

This model consists of 3 datasets, 2 of them are time series data starting from Jan 22nd 2020 until today and last dataset is constant data with country level population density information. In oder to use these datasets, 3 different input layers are added to the model for each of the dataset. These 3 input layers are then merged into one layer using Keras concatenate layer. The output of this layer is then feed into the convolutional layer. The convolutional layer is followed by stakced LSTM layers with dense and dropout regularization. The output of the model will be predicted COVID-19 cases for 159 countries. To predict cases for more days, the model needs to be fit and compiled in loop with the previously predicted data acting as an input to the new model.
One important thing about this model is that the model can have any number of time steps. Time steps determines how much historical data will be used to predict the near future data. This increased the accuracy significantly but is still not exact.

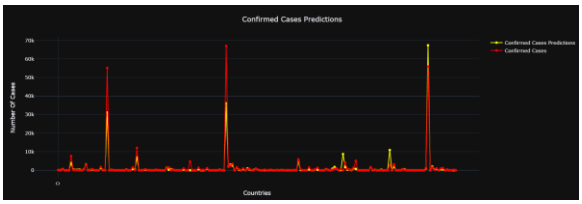
The following section talks about the results of the model in depth.

V. RESULTS

Refer the following figures to understand the output of the model.
The plots shown below are the output of the final model with training set from 22nd Jan to 31th Jul, 2020 and predicting 1st Aug to today (13th Aug10) COVID-19 cases. The plot shows predicted cases for 156 counties (in alphabetical order).

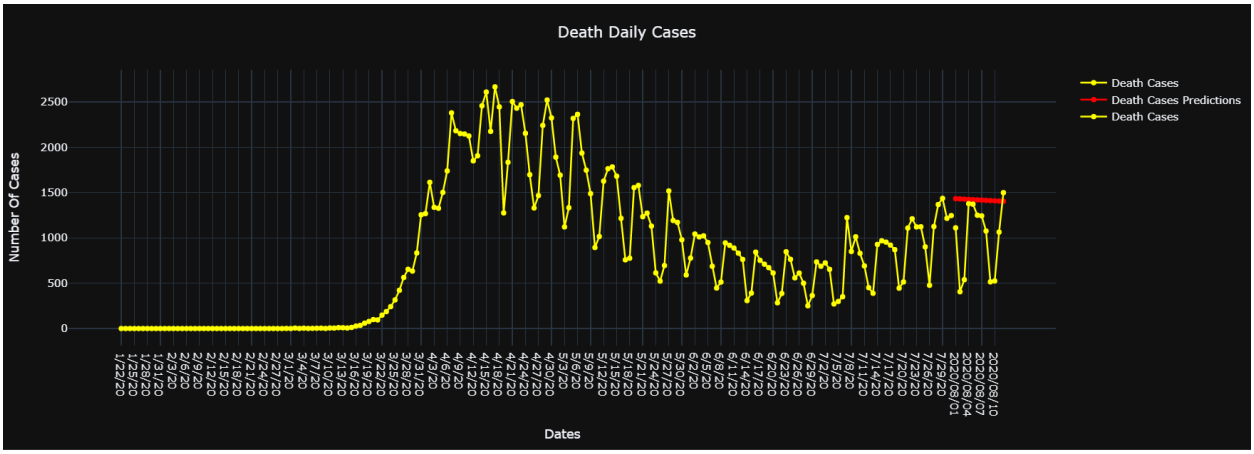


Confirmed cases for Day 1

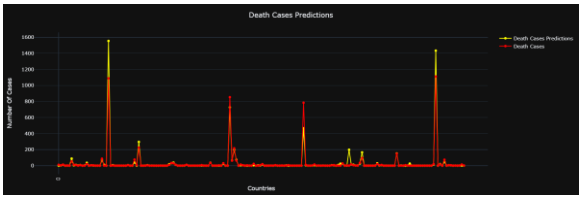


Confirmed cases for Day 13

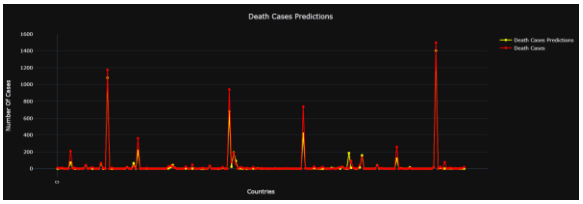
Above figures show the output of the model for confirmed daily cases. As seen above, the model does not get high accuracy for outliers (countries like US, Brazil, India). Also, the accuracy of prediction decreases as we increase the number of days since these new predictions are based on already predicted data points. Hence the error has a cumulative effect. Similar observation can be drawn for model with death case prediction.



Model Prediction for United States



Death counts for Day 1



Death counts for Day 13

Now to dig deeper and analyze model prediction for individual countries, the shown prediction output is what model generates for 'United States'.
The prediction is not accurate because United States is an outlier, but the prediction can get better with better models.

VI. CONCLUSION & FUTURE SCOPE

The modified model gives better results because we incorporate some advanced features like multi input layers, Scaling, Convolutional LSTM, Stacked LSTMs, bigger Time Steps, etc. This is a nice model to predict COVID-19 trends of any country.

In order to further work on the accuracy on the model, there are few things we can improve in the model, such as –

- Improve hyperparameter tuning, improve model architecture
- Input data regarding demographics of the countries – age, lifestyle data, medical facilities available, number of hospital beds, etc.

Further, to improve the application of the model, we can predict the stringency index for future along with the COVID-19 trends. This will help the governments understand the future policy requirement and will give them time to prepare for the next COVID-19 wave. Although, hope we get the vaccine soon.

REFERNCES

[1] John Hopkins COVID-19 Tracker, https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

[2] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." International conference on machine learning. 2013.

[3] Hochreiter, Sepp, and Jurgen Schmidhuber. ”Long short-term memory.” ¨ Neural computation 9.8 (1997): 1735-1780
[4] https://github.com/rv20197/COVID-19-Analysis-and-Prediction-Using-AI/blob/master/Cumulative%20COVID-19%20Global%20Analysis%20and%20Predictions%20.ipynb
[5] Oxford Covid-19 Government Response Tracker, https://github.com/OxCGRT/covid-policy-tracker
[6] Census data, Population density per country, https://github.com/owid/covid-19-data