Recurrent Neural Network Models for Covid-19 Confirmed Cases Prediction

Deming(Remus) Li, Yuanhao Wang, Yangkai Lin, Zhiyin Liu

Abstract

This study will present a forecast of daily new confirmed COVID-19 cases in the US based on deep learning algorithms: LSTM (Long Short-Term Memory),Bi-LSTM (Bi-directional LSTM), GRU (Gate Recurrent Unit) which are all derived from Recurrent Neural Network, and conduct comparison analysis of which model has the best prediction accuracy.

1 Introduction

The ongoing COVID-19 global pandemic has infected more than 69 million and resulted in 1.57 million deaths as of December 10th, 2020 [1]. Since the beginning of the pandemic spread in the US, there have been surges of confirmed positive cases in multiple populous states such as California, Texas, and Florida. The overall trend of state-wise infection cases continues to climb which have affected our life. This project utilizes deep learning algorithms to make an effort in predicting daily positive increase cases in the US in hopes that the methodology and the result of this project is beneficial for government and healthcare organizations to evaluate the effectiveness of currently imposed preventive measurements such as social distancing and mask wearing.

The rest of this report is outlined as follows. Section 2 describes the data pre-processing, in which we choose two data features, daily positive increase cases and recover increase cases, then we provide the visualization of them. Section 3 describes the methodology, including model algorithm, experiment setup, evaluation metrics and results. Finally, a conclusion is drawn in Section 4 to briefly discuss findings, shortcomings of this project, and future research suggestions.

2 Exploratory Data Analysis

2.1 Data Description

The datasets we intended to use are from Johns Hopkins University, which has been recording and forecasting the trend of COVID-19 cases all over the world since the pandemic started. And it is a reliable source that has been used extensively for scholars and researchers to predict all kinds of factors related to COVID-19, especially the total cases and new positive cases.

Instead of using all features to train our model, we intended to use "daily positive increase cases" and "daily recovery increase cases". The two feature selected all have significant trend that can represent the information of COVID-19. We also have visualization of those two feature to see the trend.

2.2 Data Visualization

Here is the visualization of the "daily positive increase cases" and "daily recovery increase cases":

Preprint. Under review.

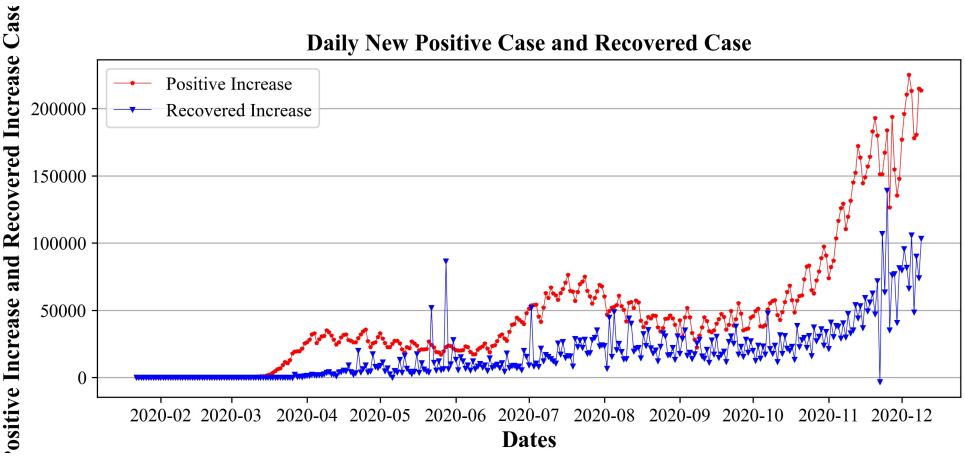

Figure 1: Positive and Recovery increase cases

3 Methods

3.1 Models

We apply three different types of Recurrent Neural Network (RNN) to our machine learning models to predict the increase of the confirmed cases in the United States. The input layer of each model receives a set of training data, and the hidden layer trains the model using the training data. Then, the model stores the changes in the hidden layer. When we input another set of training data, the hidden layer trains the model with the new set of training data and previous result. After we get the desired output, we input the testing data to predict the daily increase of confirmed cases and check the error.

3.1.1 Long Short-Term Memory

Since it can be less accurate to train a standard RNN to solve problems that require learning long-term temporal dependencies, it is mainly because the gradient of loss function decays exponentially with time. Therefore, we are going to apply the LSTM model. LSTM has a 'memory' cell that we can maintain the information for a long time period. It also has new gates, such as input and forget gates, allowing a better control the gradient flow to avoid vanishing and exploding gradient problems. In summary, the LSTM can guarantee long term dependencies.

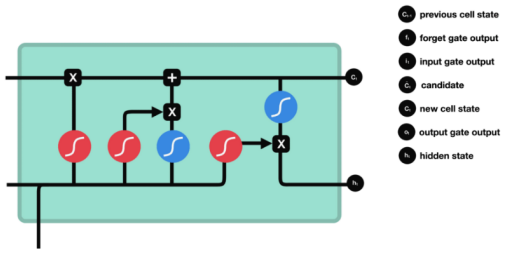

Figure 2: LSTM [1]

In LSTM, we have a cell state (store information), an input gate, a forget gate, and an output gate, where "X" in the figure above is point-wise multiplication, and "+" stands for point-wise addition. Also, sigmoid activation neurons keep the values between -1 and 1, and hyperbolic tangent activation neuron keep the values between 0 and 1.

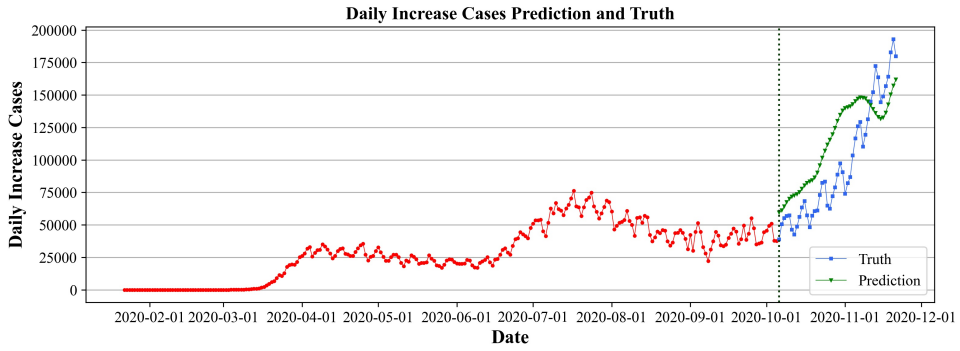

Figure 3: LSTM Testing Result Comparing With Actual Data

The figure above shows our prediction comparing with the actual data. The green line is the prediction from LSTM model, and the blue line is the actual positive increase cases recorded. The predicted trend is similar with the actual trend but with slight difference. We set each sequence with data from two features of 17 days to predict exactly the positive increase cases on 18th day. Because of the method we choose for prediction, we created a dataset of sequences for the model, and each data in the dataset contains daily increase of confirmed cases for seventeen days. When we finished our testing, we found that the number of the actual confirmed cases for each day is lower than predicted number. Then, we realize that this could be caused by some unavoidable factors, such as raining, which prevents people going to hospitals. Thus, it is reasonable for the actual data to be lower than our prediction.

3.1.2 Bidirectional Long Short-Term Memory

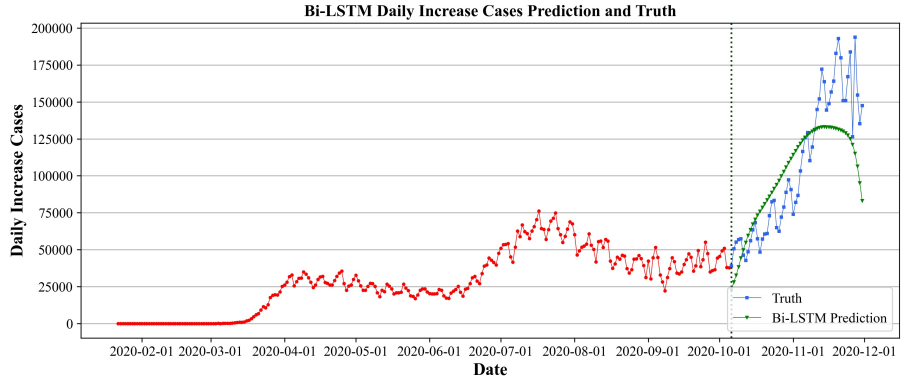

Figure 4: Bi-LSTM Testing Result Comparing With Actual Data

For the trend comparison, the prediction from Bi-LSTM seems smoother than other two models. This is because there are two directions of data: one from past to future and one from future to past. The reason that it drops near the end, we concluded, is that there is a lack of data from future to past, which is reasonable according to our implementation and data structure. It is still a great model to predict middle values given its stability. We believe that Bi-LSTM is more suitable for NLP such as predicting words within a sentence. We will finally choose the most appropriate future prediction in our report. Therefore, the 45-days prediction is not included in this report based on our consideration, and we will focus on analyzing LSTM and GRU models.

3.1.3 Gate Recurrent Unit

In GRU, we have an update gate that decides what new information to add and what throws away with an input, a reset gate that is used to decide how much past information to keep or forget, and an

output gate. Point-wise addition, point-wise multiplication, sigmoid function, and hyperbolic tangent function are same as what we have for LSTM.

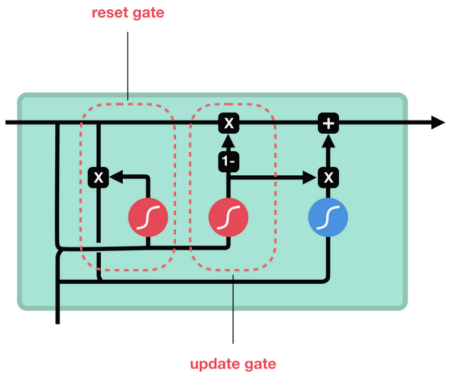

Figure 5: GRU [1]

However, LSTM and GRU also have some differences. LSTM more focuses on current data. In other words, data from September to October in our dataset. And GRU focuses on the whole dataset.

The figure below shows the testing result of GRU. Since GRU is similar with LSTM, we use the same method as for LSTM to test our model. The predicted trend is also similar with the trend predicted by LSTM.

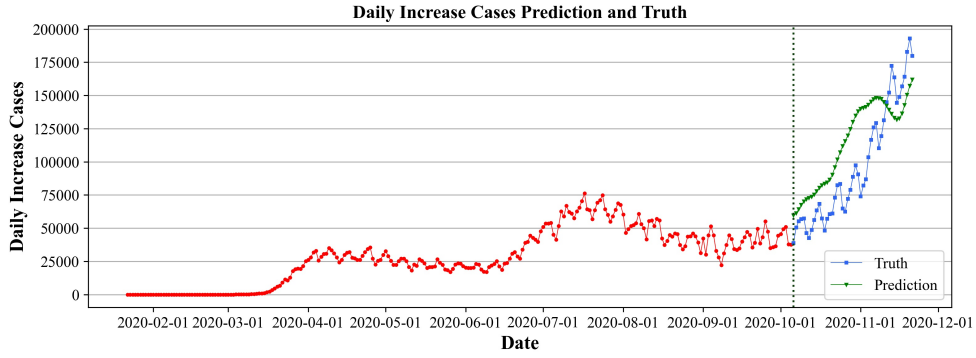

Figure 6: GRU Testing Result Comparing With Actual Data

3.2 Experimental Settings and Evaluation Metrics

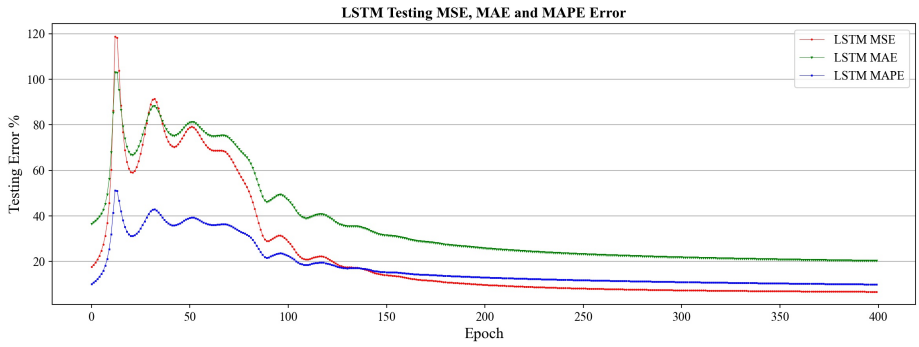

Figure 7: LSTM Error

Each prediction model is implemented using built-in PyTorch deep learning APIs. The dataset is firstly normalized to share the same ranges of values for the different features. 80% of the data is used for training which is approximately 261 days, and 20% of the data is used for testing. For the training phase, Adam optimizer is employed with a learning rate of 0.001, and the total number of training iterations is 350. MSE (Mean Square Error) loss function is used to optimize each prediction model. For each iteration, MSE, MAPE (Mean Absolute Percentage Error), and MAE (Mean Absolute Error) are calculated. By conducting the experiment on each model with similar settings, we analyze the prediction performance of these models by comparing their overall testing errors.

3.3 Experimental Results and Discussions

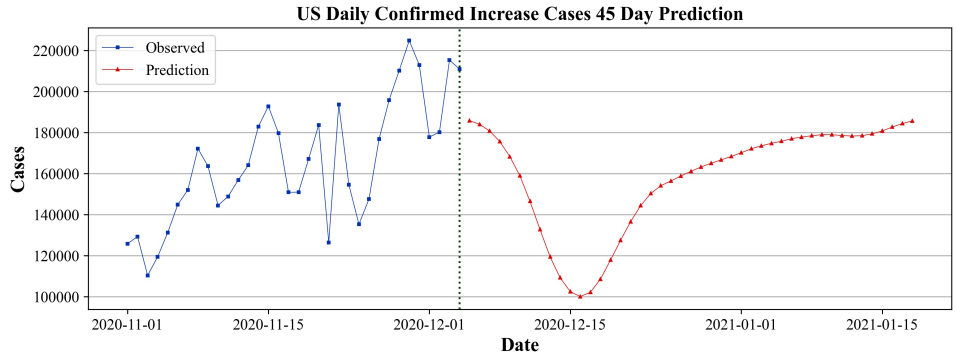

Figure 8: Model Prediction

GRU and LSTM generate similar results when we apply the newest data from Johns Hopkins University to both models. The figure above shows the result. From the figure, we can learn that the positive increase cases start decreasing after December 1st and increasing around December 15th. One reason for this change, we guess, could be that students in most schools in the United States will be taking final exams during that period. Therefore, most families will not go travelling, which can prevent the spread of the virus. After December 15th, students have finished their final exams, and they probably will plan on vacation with their family or friends. Population movement is the main cause of the massive spread of the virus.

4 Conclusion

To conclude, we trained LSTM, Bi-LSTM, and GRU, which are three different models derived from Recurrent Neural Network, by setting 17 sequences to predicted the 18th day's positive increase cases. We analyzed that both GRU and LSTM have good predictions during the training part on time series dataset. However, since the lack of data from future to the past, it is reasonable that Bi-LSTM can not generate a good prediction according to our dataset. But it is still a great model to study on in the future.

Base on our prediction of 45 days in the future, the trend is beneficial for government and healthcare organizations to evaluate the effectiveness of currently imposed preventive measurements such as social distancing and mask wearing. Also, citizen can better plan on future life base on the trend.

In the end, our team sincerely hopes everyone stay healthy and safe. Do not forget to wear a facial cover in public area. Be well! Be committed! We all work together to go through this pandemic!

5 Reference

[1] Medium. 2020. Illustrated Guide To LSTM'S And GRU'S: A Step By Step Explanation. [online] Available at: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> [Accessed 23 October 2020].

[2] Valkov, V., 2020. LSTM Time Series Prediction Tutorial Using Pytorch In Python | Coronavirus Daily Cases Forecasting. [online] Available at: <https://youtu.be/8A6TEjG2DNw> [Accessed 23 October 2020].

6 Contribution

Zhiyin Liu:

1. Data visualization: The trend of positive cases of Top 10 states.
2. LSTM model building model training and model prediction
3. Editing report together
4. Worked on the presentation.

Deming Li:

1. Data visualization: The trend of positive cases.
2. Data pre-processing, LSTM, Bi-LSTM model building, model training and model prediction
3. Report writing: Figure explanatiosn, conclusion

Yuanhao Wang:

1. In charge of RNN related work.
2. Edit the Latex format.
3. Worked on reports with teammates.
4. Worked on the presentation.

Yangkai Lin:

1. Data visualization: Positive increase and recovered increase cases
2. Data pre-processing, model building, model training and model prediction
3. Report writing: abstract, introduction and experiment result discussion.