Analysis and Prediction of Trends in the Covid-19 Virus Using LSTM Between Countries of Different Population Density

Toby (Lingxuan) Chang
20336690

Jinyang Huang
20812495

Guanying Zhao
20829577

CS 680 Final Project
April 7th, 2020

Abstract

In this report, we investigated future trends of the novel virus, Coronavirus (Covid-19), using the Long short-term memory (LSTM) recurrent neural network (RNN). We compared and contrasted performance in minimizing loss between different combinations of loss functions and optimization algorithms used. We also analyzed the accuracy of predictions for the LSTM model when trained from data of individual countries versus aggregation of data from all countries. Lastly, we used the trained model to predict the future trajectory of the virus and compared between different countries, which have very different demographic and population characteristics. We also want to investigate the effectiveness of social distancing protocols in curbing the spread of the disease by using a country's population density as an indirect measure of how well social distancing protocols are practiced by citizens of a country.

1 Introduction

1.1 Data Processing

The time series data of the Covid-19 virus was obtained from the Github repository provided by the Johns Hopkins University Center for Systems Science and Engineering who having been collating global data since the beginning February of this year (3). The data is updated once everyday around midnight coordinated universal time (3). The specific datasets used included:

- confirmed positive test cases of the Covid-19 infection, which included cases that were confirmed both in either laboratory or clinical setting (3).
- cases of death due to the virus (3)
- cases of recovered from the virus (3)

The time series dataset contains a sequence of case counts that are 24 hours apart and has the fields: province, country, latitiude, longitude, and the date fields with the count of confirmed cases, deaths, or recovery for that day (3). Some minor cleaning was performed where we drop the province, latitude, and longitude fields as well as removed some countries that had very few number of confirmed cases. We also merged the counts for each date for each province that belonged to the same country and summed them up.

We also obtained data for population density of countries from Wikipedia. We only used the fields: country name and density. The unit used for density was population per square kilometre ($population/km^2$), which is the number of humans that are located in a certain area per square kilometre. This characteristic is of great importance since many countries, starting with China, have begun to adopt practices of social isolation and **social distancing** in hopes of slowing down the spread of the virus. The aim is to flatten the curve of the virus infection. Instead of having a large amount of individuals getting infected in a short amount of time, we have the same amount of individuals infected but over a much longer period of time. This prevents a country's health care system from being overburdened with a sudden increase in patients and a demand for medical resources. Since the virus spreads mostly by human-to-human contact, social distancing protocols may be effective in curbing the spread of the virus because we are physically distancing ourselves from other people, and thus reducing the likelihood of contact and transmission when sneezing or coughing.

We chose population density as an indirect measure of how effective social distancing can be implemented and enforced in each country since population density directly reflects the amount of individuals contained in an area of space. To illustrate, if the density of an area is very high, then the distance between individuals when spread out evenly from each other will be very small compared to a scenario where the density of an area is very low. Thus, we can infer that a country with a high population density will have a much harder time of carrying out and enforcing social distancing protocols, and thus slowing down the spread of the pandemic, compared to countries with a low population density.

Finally, we took the count of cases for each country for each date and divided over that country's respective population density. The resultant value indicates the number of confirmed cases per square kilometre. We then normalized this value between 0 and 1.

2 LSTM

We chose the LSTM model because it is a time series data. It is a unique kind of recurrent neural network that has the performance capabilities to solve the problem of long term dependencies. For instance, we need the count of confirmed cases from January 30th in order to get an accurate prediction for count of cases in April 15. This growing gap between January 30th and April 15th is the long term dependencies that prevents regular RNNs from connecting the information. For each cell state in the RNN sequence, the cell takes the count of cases from 4 consecutive dates and outputs the prediction for subsequent 5th date. The model was trained using Keras which had better performance than Pytorch. Tensorflow 1.15 was used and runs were performed on Google Colab.

3 Problem Definition

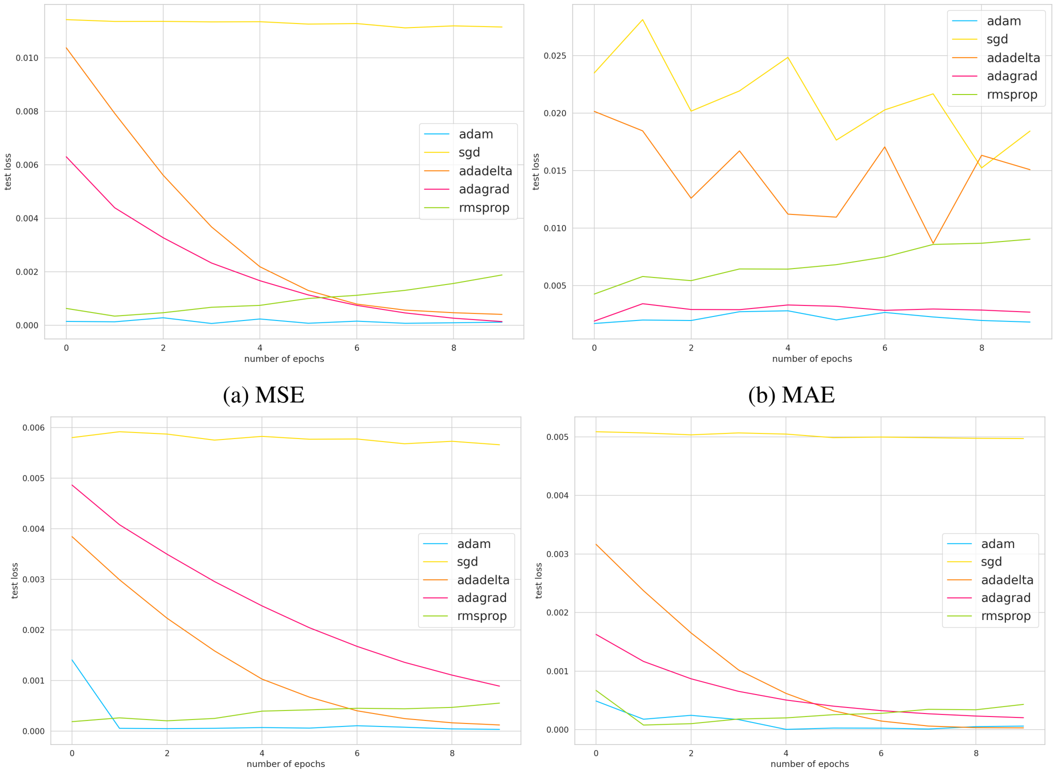
The objective of our project includes:

- Applying the five different optimization algorithms: Adam, Adadelta, Adagrad, RMS, and Stochastic gradient descent in combination with each of the four loss functions: Hinge, Mean Squared Error (MSE), and Mean Absolute Error (MAE), and Log Cosh to LSTM and compared the performance in minimizing the loss
- Compare how well the LSTM model performs when trained from data of individual countries versus data aggregated from all countries
- Compare future trajectory of Covid-19 between different countries and looking at each respective country's population density, infer the effectiveness of social distancing protocols in that country.

4 Discussion

We selected Mean squared error, Mean absolute error, Huber loss and Log cosh loss functions combined with each of the following optimization algorithms: SGD, Adam, Adagrad, Adadelta and RMSProp. We compared each of the above twenty combinations' performance in minimizing loss.

As shown in figure 1, the Adam optimizer for every loss function performed the best and converged the fastest. MAE was the worst loss function where every optimizer did not converge near zero. SGD was the worst performing optimizer in each combination of loss function. The reason might have been that SGD is not adaptive for learning rate and does not set a momentum. The constant learning



(a) MSE

(b) MAE

(c) Huber Loss

(d) Log Cosh Loss

Figure 1: Comparison of testing performance among different loss functions and optimizers in 10 epochs on all data

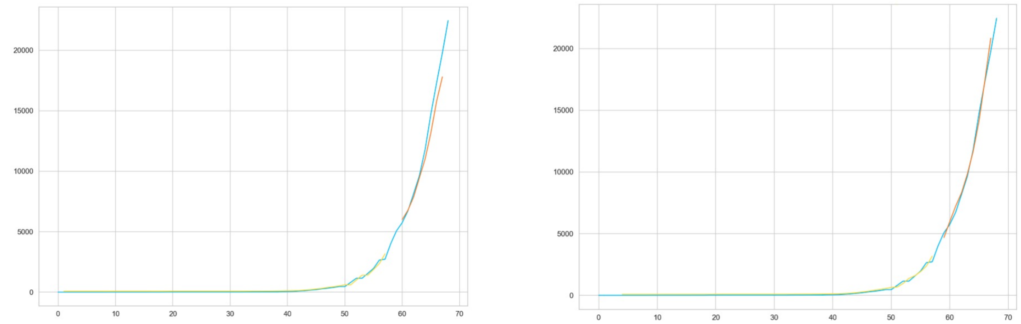
rate that was used is not large enough. Thus, it always bounces around local minima. Although, Adagrad, Adadelta and RMSProp seems good in some specific cases, we chose the Adam optimizer combined with the MSE loss function for the rest of our experiment since it had the most consistent converge rate performance.

We also compared how the model performs when trained for different epochs and look back values. We found out that running for 70 epochs and using a smaller value of 7 days for the look back parameter produced the best results. This is probably because running for too little or too many epochs results in under training and over-training respectively, which results in underfitting and overfitting of the data. A smaller look back value was associated with better results. This is most likely because a smaller look back value results in more permutations of input vectors for training compared to larger look back values. For instances, 5 days can be split into 3 input vectors of size 3: $[1, 2, 3]$, $[2, 3, 4]$, and $[3, 4, 5]$ and only 2 input vectors of size 4: $[1, 2, 3, 4]$ and $[2, 3, 4, 5]$.

To explore the performance of LSTM, we trained the model using data of individual countries and also data aggregated from all countries. Dataset was split between training and testing set. We set LSTM to look back 7 days and ran for 70 epochs for each case. Then, we predicted for both training and testing datasets in both cases and compared with the true cases. We hypothesis that the model trained from the aggregation of data of all countries will produce more accurate predictions than the model trained from data of individual countries. This is because taking all countries' data into account provides more information in several ways.

First, since the spread of the virus began at different times in each country, some countries such as China has already overcome the pandemic while countries like the US is still suffering from increased daily cases. Thus, the model trained from all data would be able to infer how the trend would go for countries like US by learning from the curves of those countries that has already been at that stage of the pandemic. The Model trained on a single country's data would probably never flatten since it only sees the increasing trend for its own current time period.

Second, using the aggregation of data of all countries avoids overfitting, which would have most likely happened if we only used a single country's data.

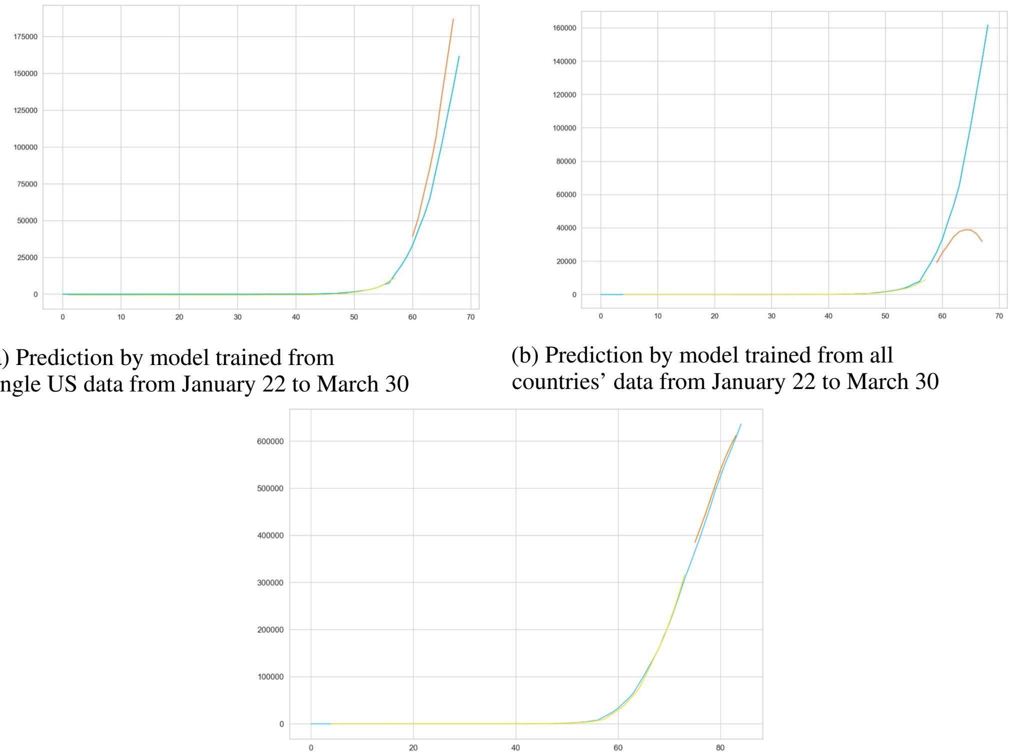


(a) Prediction by model trained from single UK data from January 22 to March 30

(b) Prediction by model trained from all countries' data from January 22 to March 30

Figure 2: Comparison of Prediction performance between model trained on single country's data versus all countries' data (UK)

For instances, Italy in comparison to China has a very different starting day for the pandemic as well as a very different population characteristics and density. Italy also has very different methods of enforcing social distancing protocols in contrast to China. Thus, if the model trained on only Italy's data is used to predict for other countries like China, it will result in a lot of sampling and selection bias due to overfitting. However, the model trained from all countries' data will minimize this bias.



(a) Prediction by model trained from single US data from January 22 to March 30

(b) Prediction by model trained from all countries' data from January 22 to March 30

(c) Prediction by model trained from all countries' data from January 22 to April 15

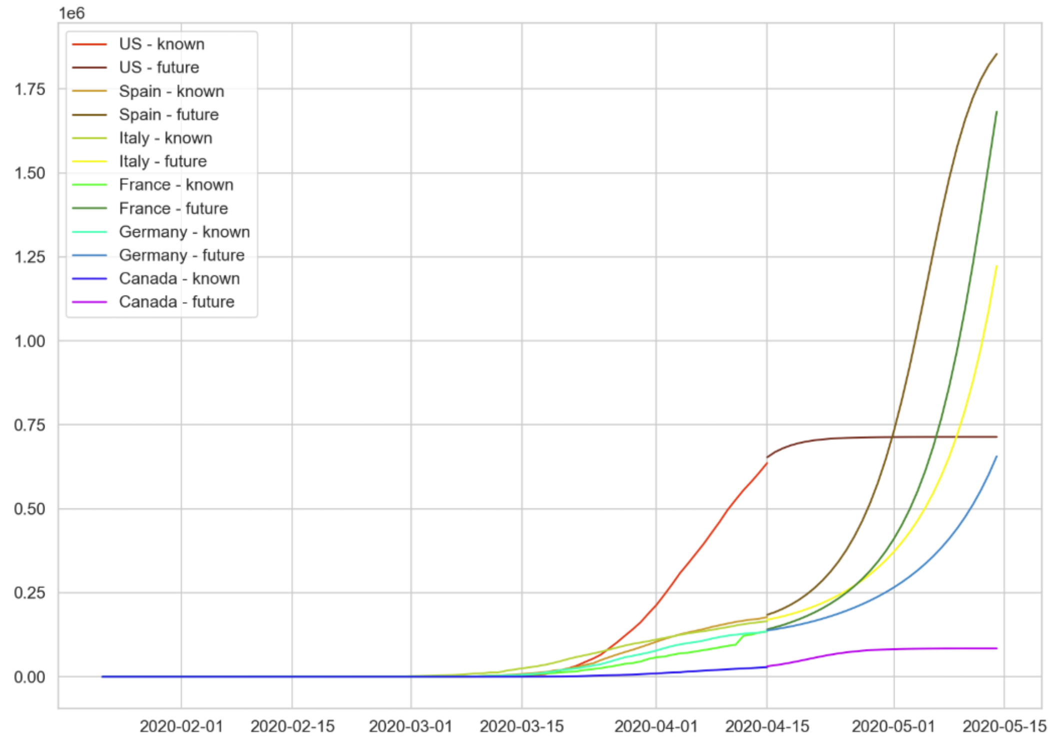
Figure 3: Another Comparison of Prediction performance between model trained on single country's data versus all countries' data for data between January 22 to March 30 versus data between January 22 to April 15 (US)

As depicted in figure 2, we can see that the model trained from all countries' data in figure 2b has better performance with prediction with a much smaller testing score of **458.21 RMSE** than the model trained from a single country's data in figure 2a with a testing score of **704.26 RMSE**. The yellow and orange curves mirror very closely the blue curve. **The blue curve shows the true cases. The yellow line shows the training predictions and the orange line shows the testing predictions**.

However, there are some drawbacks. We ran the model on two versions of the data. One contains confirmed cases from January 22nd to March 30th (smaller dataset). The other contains cases from January 22nd to April 15th (larger dataset). This model does not perform very well for some countries when the data is relatively small, such as the US. This is depicted in figure 3b where the prediction for the last 10 days shown in orange has started to flatten. The testing score had a value of **54753.36 RMSE**, which is extremely large. This may be because China is the only country that has reached the peak of the curve while all other countries are still on there way to the peak up till March 30th. Thus, the model may have been biased by China's trend. And we can see that when US is trained sololy using its own data, the testing score is **4406.40 RMSE**, which is much smaller.

In contrast, when we use the newest data (January 22nd to April 15th), the model performs much better as seen in figure 3c for the US. This may be because many countries have started to develop a similar trend as China by reaching further up the curve or starting to flatten. The prediction performance is still better when using all countries' data versus single country's data when comparing figure 3a and figure 3c. The testing score also decreased to **47256.27 RMSE**.

For the analysis of future predictions, we trained the model with data from all countries with the newest data (January 22 to April 15). Then, taking the true cases for the last number of days up till April 15 depending on what value the parameter look back was set to, we looped for 30 iterations with each iteration predicting the cases for subsequent days after April 15. For instances, when look back is set to seven days, we start by taking the true cases from April 9 to April 15 as the first input vector. We feed this vector into the model and it predicts the cases for April 16. Then, we appended the predicted case for April 16 to the initial input vector and we took the cases for the new last seven days from April 10 to April 16 as the new input vector.



(a) Future Prediction for April 16 to May 15
by model trained for 70 epoch

Figure 4: Future Predictions made by model trained for 70 epochs

Figure 4 shows the current known cases from January 22nd to April 15 as well as the future predicted cases from April 16 to May 15 for the top five countries with the most confirmed cases according to Wordometer. We also plotted the trend for Canada as well. We see the the curves for US and Canada flattening out at 750,000 and 100,000 cases while Spain, Italy, and France continue to rise with Spain starting to flatten at close to 1.75 million. The population density for these six countries from highest to lowest is: Germany ($233pop./km^2$), Italy ($200pop./km^2$), France ($123pop./km^2$), Spain ($93pop./km^2$), US ($34pop./km^2$), Canada ($4pop./km^2$). We can see that the prediction for the 4 European countries with the larger population densities show very steep curves while the US and Canada with the smaller population densities show shorter curves that is starting to flatten out.

5 Conclusion and Limitations

All in all, we discovered that our LSTM model has the best performance when using a combination of MSE and Adam as the loss function and optimizer respectively while running for 70 epochs with a look back value of 7 days. Although there does seem to be some indication of high population density related to higher future predicted cases, we cannot solely infer this correlation using population density alone. For instance, China has a population density of $145pop./km^2$, but China has already controlled the spread of the virus and flattened the curve. Thus, there are many other factors at play that we have not considered in our model, such as how the virus spreads, population characteristics such as air pollution (which may accelerate the spread of the virus), each country's policy in enforcing social distancing and quarantines.

In addition, our data solely relies on clinical and laboratory diagnoses positive test cases. Countries may differ in detection rate of the virus because certain individuals may not voluntarily report symptoms and get tested. Countries also differ in number and quality of medical resources such as testing kits and medical professionals to conduct the tests. Hence, countries that detect more will have more confirmed cases. Furthermore, some individuals may have already died before being tested for the virus. It is impossible to know the true numbers. The data is significantly impacted by sampling and selection bias.

References

[1] Brownlee, Jason. "Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras." Machine Learning Mastery, 5 Aug. 2019, machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[2] "Coronavirus Cases:" Worldometer, www.worldometers.info/coronavirus/

[3] CSSEGISandData. "CSSEGISandData/COVID-19." GitHub, 15 Apr. 2020, `github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data`

[4] "List of Countries and Dependencies by Population Density." Wikipedia, Wikimedia Foundation, 12 Apr. 2020, `en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density`