CS5891 Final Project: COVID-19 Prediction

Yue Hu

Abstract

Since the onset of COVID-19 around Feb. 2020, millions of people all around the world are affected. In this work we build an LSTM model to predict the number of confirmed cases for each country. Dataset provided by Kaggle is used, summarizing global infection cases up till April 7. Different hyper-parameters are tested to find the best fitting model. The result shows an accurate prediction for most of the countries except for the extreme case of US.

1 Introduction

Since the onset of COVID-19 pandemic around Feb. 2020, the life of every one all around the world is influenced. As of now, we are still performing social distancing, wondering how long shelter-at-home would last. A model predicting future confirmed cases might be helpful to answer such questions.

To better understand the pandemic, Kaggle is providing a dataset recording the confirmed cases all around world [1], which is the effort of research groups and companies organized by The White House Office of Science and Technology Policy (OSTP). While the original competition aims at addressing a series of key open scientific questions on COVID-19 [1], here we only focus on a specific question: given the historic data, can we accurately forecast the number of confirmed cases for each country? Building such model could be the first step towards understanding the spreading pattern of COVID-19, and the resulting prediction could aid our decision making for various policies.

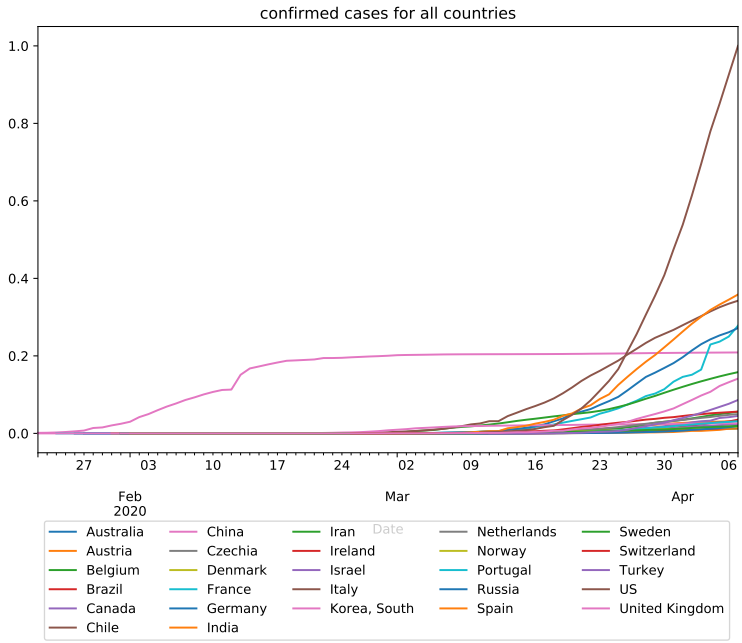

Figure 1: Confirmed cases for all countries.

The rest of the article goes as following. Section 2 formalizes the problem. Section 3 explains the approach and interpreters the results on a high level. Then Section 4 explains the technical details. Section 5 discusses related literature. Section 6 states limitation and future works.

2 Problem statement

The problem to be solved is as follows. We aim to predict the number of confirmed case of a specific country given its historic records of confirmed case. More specifically, we predict the number based on the history of past seven days. We assume that each country follows a similar patter. That is, every country can be predicted by the same LSTM model. This assumption can be justified from Fig 1, where every country roughly follows an exponential initial curve, except for China whose number becomes steady after a month.

The fitness of the model is measured by how well it predicts the confirmed cases of an unseen date. Mean squared error (MSE) is used as criterion.

3 Approach

This section briefly explains our approach, and qualitatively interpreters the outcome on a high level.

This prediction task is a typical time series prediction problem. LSTM is used, as it is one of the most widely used DL model for sequential data. Details of the model is explianed in Section 4.

The overall result is shown in Fig 2, and Fig 3 zooms in for each country. The dotted grey line is ground truth data, and the colored lines are predictions. We can see they matches quite well. In fact, the lines almost covered the grey dots in most cases. This shows the effectiveness of LSTM in our task of predicting COVID-19 time series.

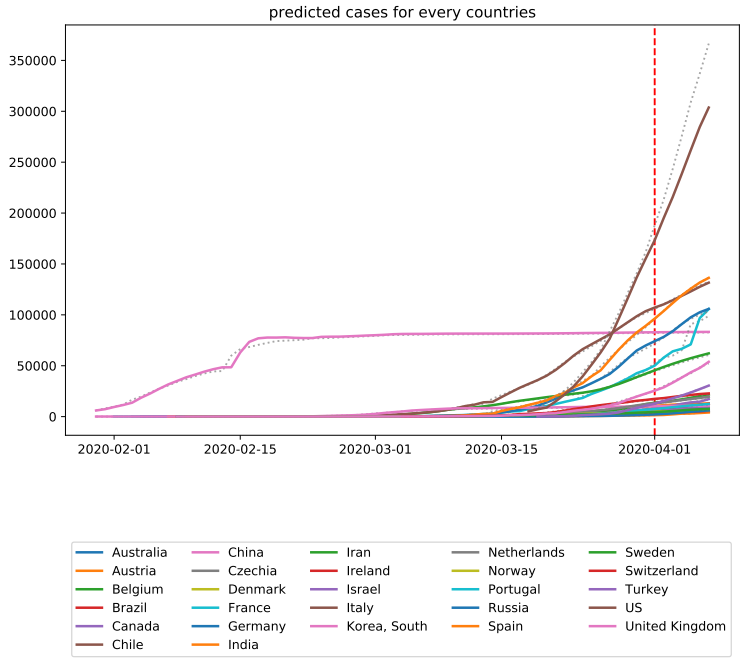

Figure 2: Predicted cases for all countries. The dotted grey lines are ground truth data, and the colored lines are predictions. On left side of vertical red line is training data, and on the right unseen testing data. In fact the lines almost covered the grey dots, showing exact prediction.

There are two countries where prediction noticeably deviates from the ground truth, namely US and France. For US, the model believes that the pandemic should spread at a lower speed beginning April 1. This may be because that cases in US is so large and grows at a higher rate than the rest countries, so

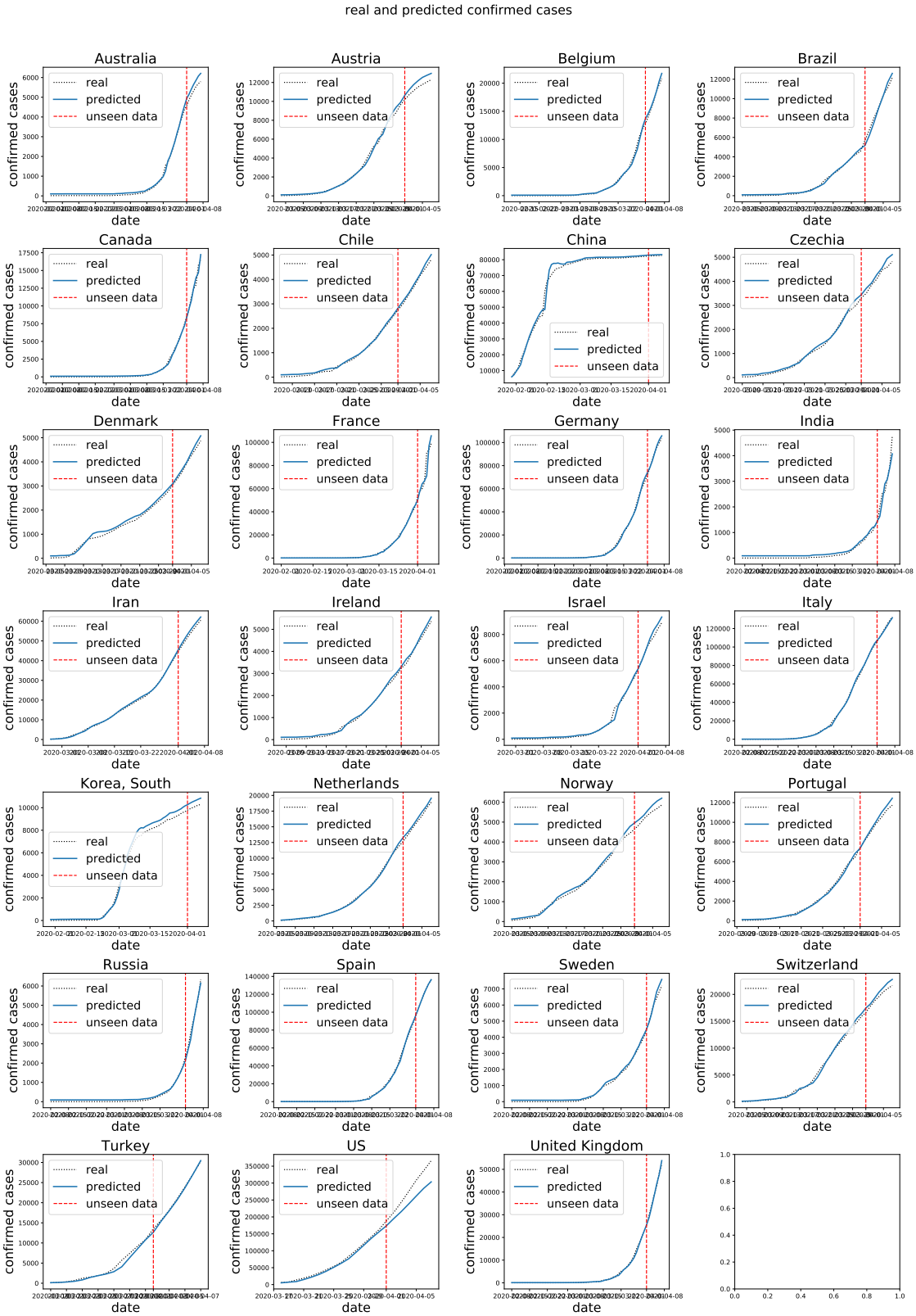

Figure 3: Zoomed-in prediction for each country. The dotted grey lines are ground truth data, and the blue lines are predictions. On left side of vertical red line is training data, and on the right unseen testing data. In fact the lines almost covered the grey dots, showing exact prediction.

learning the history of other countries cannot help with predicting such drastic increase in US. For France, there is a sudden increase in slope around April 2, which causes some error in the model.

4 Model Details

In this section, the technical details of the LSTM model are explained.

4.1 Data processing

The raw data contains confirmed cases time series for each country. Some data cleaning and pre-processing is done. Firstly only the major infected countries are selected. The threshhold is chosen to be a minimum of 5000 cases reached. Then each country starts to count from the occurrence of its first case. That is, all 0 confirmed cases entries are thrown.

A minmax scaler is used to scale all numbers between 0 and 1. Then, a sliding window is used to process the data into trainable format. For each country each date, input data X is the time series of previous 7 days, and label y is the confirmed case of that day. In reality, this mimics the process of prediction next day's confirmed case based on the information of the previous week. Slices of all countries are mixed together.

Cases before April 1 are set as training data, and after April 1 testing data. The resulting training inputs X has shape (1096, 7), training labels has shape (1096,). Testing inputs X has shape (162, 7), testing labels has shape (162,). (Actually they should be called validation data and are used for hyper-parameter-tuning. Kaggle has a separate un-revealed test data for "real" prediction.)

4.2 LSTM model

A standard LSTM is used. The LSTM is set to have single layer, with input size 1. The hidden size is varied to find the best model structure hyper-parameter. A fully connected layer is used in the end to connect the hidden layer and output a single prediction value. The results for various hidden size is shown in table 1. We can see that as size of hidden layer go up, error first goes down then up. Best hidden size is 5.

Table 1: Result of MES loss for different LSTM hidden sizes.

hidden size	2	5	10
training loss	1.141578e-05	3.575298e-06	3.603011e-06
testing loss	8.869912e-04	**2.995182e-05**	3.174603e-05

4.3 Tuning of hyper-parameters

Firstly, the impact of learning rate is investigated, as shown in table 2. We can see that learning rate 0.01 is best for our LSTM model

Table 2: Result of MES loss for different learning rates.

learing rate	0.001	0.005	0.01
training loss	1.107143e-04	3.497536e-05	4.879545e-06
testing loss	1.816602e-03	1.970778e-03	**3.825084e-05**

Secondly, the impact of L2 norm regularization is investigated, as shown in table 3. We can see that a moderate l2 normolization (1e-6) is helping in this case.

Lastly, the impact of optimization method is investigated, as shown in table 4. We can see that Adam performs much better than Gradient descent.

Table 3: Result of MES loss for different learning rates.

l2 regularization	1e-8	1e-6	1e-4
training loss	3.846330e-06	4.032986e-06	7.353704e-05
testing loss	7.029613e-05	**6.478220e-05**	2.067197e-03

Table 4: Result of MES loss for different learning rates.

optimization method	ADAM	Gradient descent
training loss	4.032986e-06	3.156327e-03
testing loss	**6.478220e-05**	3.078640e-02

5 Related work

Long short-term memory (LSTM)[2], since its introduction in the 20th century, has been widely used for time series models. Its ability to incorporate in both long and short term dependencies has enabled it to learn many complicated real world processes. LSTM has been successfully applied to many fields of studies, such as traffic [3], natural language processing [4], sensor network [5], etc.

In this article, we apply LSTM on COVID-19 spread prediction. Prior to our work, many physics dynamic model has been applied to learn the spread of COVID-19, such as [6, 7, 8, 9]. Early on, deep learning models are applied to study the transmission dynamics of the epidemics in China [10, 11]. Here we apply LSTM to a wider region, and study its spread globally.

6 Conclusions

In this work we build an LSTM model to predict the number of confirmed cases for each country. Different hyper-parameters are tested to find the best fitting model. The result shows an accurate prediction for most of the countries except for the extreme case of US.

Regarding future work. This dataset is up till mid April. Now as the situation evolves, we might be able to better gather the data, retrain the model and make predictions up to date.

Another limitation is the interpretability of the model. Ideally, we want a model that can not only predict precisely, but can demonstrate what are the factors that slows down or speeds up the spread of COVID-19. Further work would be meaningful which incorporates in various factors such as weather an policies and analysis the sensitivity of the factors.

References

[1] kaggle-covid19 global forecasting. `https://www.kaggle.com/c/covid19-global-forecasting-week-4`, 2020. Accessed: 2020-4.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197, 2015.

[4] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*, 2016.

[5] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

[6] Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic analysis of covid-19 in china by dynamical modeling. *arXiv preprint arXiv:2002.06563*, 2020.

[7] Zhihua Liu, Pierre Magal, Ousmane Seydi, and Glenn Webb. Predicting the cumulative number of cases for the covid-19 epidemic in china from early data. *arXiv preprint arXiv:2002.12298*, 2020.

[8] Gábor Vattay. Predicting the ultimate outcome of the covid-19 outbreak in italy. *arXiv preprint arXiv:2003.07912*, 2020.

[9] K Roosa, Y Lee, R Luo, A Kirpich, R Rothenberg, JM Hyman, P Yan, and G Chowell. Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020. *Infectious Disease Modelling*, 5:256–263, 2020.

[10] Zixin Hu, Qiyang Ge, Li Jin, and Momiao Xiong. Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112*, 2020.

[11] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3):165, 2020.