Research on the Forecast of the Spread of COVID-19

LIHAO GUO

James E Rogers College of Law, The University of Arizona, Tucson, 85719, Arizona, USA;
leolihao@email.arizona.edu

YUXIN YANG

James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK

With the spreading of COVID-19, various existing machine learning frameworks can be adopted to effectively control the epidemic to help research and predict the spread of the virus before the large-scale application of vaccines. Based on the spatiotemporal graph neural network and mobility data, this paper attempts to offer a novel prediction by building a high-resolution graph with the characteristics such as willingness to wear masks, daily infection, and daily death. This model is different from the time series prediction model. The method learns from the multivariate spatiotemporal graph, the nodes represent the region with daily confirmed cases and death, and edges represent the inter-regional contacts based on mobility. Simultaneously, the transmission model is built by a time margin as the characteristic of the time change. This paper builds the COVID-19 model by using STGNNs and tries to predict and verify the virus's infection. Finally, the model has an absolute Pearson Correlation of 0.9735, far from the expected value of 0.998. The predicted value on the first and second day is close to the real situation, while the value gradually deviates from the actual situation after the second day. It still shows that the graph neural network uses much temporal and spatial information to enable the model to learn complex dynamics. In the future, the model can be improved by tuning hyper-parameter such as modulation numbers of convolution, or construction of graphs that suitable for smaller individuals such as institutions, buildings, and houses, as well as assigning more features to each node. This experiment demonstrates the powerful combination of deep learning and graph neural networks to study the spread and evolution of COVID-19.

CCS CONCEPTS • Computing methodologies~Machine learning~Machine learning approaches~Neural networks

Additional Keywords and Phrases: Covid-19, GNN, STGNNs, forecast, high-resolution movement, wearing masks

ACM Reference Format:

Lihao Guo and Yuxin Yang. 2020. Research on the Forecast of the Spread of COVID-19. In 2021 11th International Conference on Biomedical Engineering and Technology (ICBET 2021), March 17-20, 2021, Tokyo, Japan.. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3460238.3460246

1 INTRODUCTION

The ongoing deadliest pandemic of COVID-19 affects the world, which has infected over 75,110,651 cases and caused 1,680,395 deaths worldwide as of December 20th, 2020 [1]. Besides, it has had an unprecedented impact on the global economy and people's ordinary lives.

This paper will analyze the training and modeling by establishing the spatiotemporal neural network and the conferral of social characteristics on the model based on non-mandatory suggestions such as wearing a mask and keeping a social distance. It will also predict the spread of the virus and analyze the virus's spread under social conditions.

In the context of the COVID-19 pandemic, various researchers worldwide study the spreading model and prediction of the new coronavirus. There are two primary models of viral transmission. The first is the traditional mechanical approach, which simulates and predicts disease spread in a predefined set of individuals or groups

[2,3]. The second method is time-series learning, such as autoregression, curve fitting, and deep learning of time series data [4,5,6]. These methods usually assume a relatively closed system that does not explicitly indicate a propagation location or a given location's prediction only depends on information from that location. With researchers' continuous efforts globally, there is a newer mechanical transmission model of spatiotemporal resolution based on regional interventions and patient contact tracing [7]. There is also a spatiotemporal graph neural network model with the mobility of personnel characteristics under the time-series learning method [8]. However, there are too many factors affecting virus transmission in reality, and the existing frameworks and models cannot correctly reflect virus transmission. This study will forecast the current Coronavirus transmission in the United States by combining traditional mechanical and time-series learning methods. According to the traditional communication model, this study redefines message passing from the neural network by adding behavioral characteristics, such as willingness to wear a mask, in order to carry out high-resolution spatiotemporal graph neural network modeling and prediction.

2 METHODOLOGY

2.1 Graph Neural Networks

In recent years, there has been increasing interest in graphical extensions of deep learning methods. Data is derived from non-Euclidean domains and represented as graphs of complex associations between objects. Recent improvements in deep learning techniques have predicted that the intricate spatial-temporal crime patterns are more tractable. At the same time, Graph Neural Networks (GNNs) came into existence. GNN is a connection model that captures graph dependencies through message passing between graph nodes. Unlike standard neural networks, graph neural networks retain a state in which any deep neighborhood represents information . Subsequently, a neural network classification method that divides advanced satellite systems into four categories is proposed, namely recurrent graph neural network (RNN), convolution graph neural network (CNN) , graph autoencoder and Spatio-temporal graph neural network (STGNN) [9, 10]. The motivations of GNNs roots in convolutional neural networks (CNNs), as well as graph embedding, which learns to represent graph nodes, edges, or sub-graphs in low-dimensional vectors. GNNs are based on CNNs and proposed graph embedding to aggregate the information in the graph structure. Therefore, they can model inputs and outputs composed of elements and their dependencies. Furthermore, RNN can be used to simulate the diffusion process simultaneously on the graph.

2.2 Spatio-Temporal Graph Neural Network

CNNs have become an increasingly active researching field, which uses predefined Laplacian matrices based on node distances to model nodes' spatial dependencies in graphs. In contrast, the core of adopting STGNNs is to examine spatial dependency and temporal dependency of the data simultaneously [11]. STGNNs aim to learn hidden patterns from spatial-temporal graphs, which become increasingly important in various traffic forecasting applications [11]. Many existing methods capture spatial dependencies by integrating graph convolutions and use RNNs or CNNs to model time dependencies. There are optimistically various advantages of applying spatial features in this work. Graphs are natural representations for a wide variety of real-life data in social, biological, financial, and many other fields. GNNs-based deep learning methods have recently shown superior performance on several tasks, including semi-supervised node classification link prediction, community detection, graph classification, and recommendations. Spatio-temporal graph is a kind of graph that models' connections between nodes as a function of time and space and has found uses in a wide variety of fields. GNNs successfully applied to Spatio-temporal traffic graphs and (especially relevant to this work) Spatio-temporal influenza forecasting. In these latter two cases, both primarily incorporated temporal dependencies at the model level, either through the decomposition of a dynamic Laplacian matrix or through a recurrent neural net.

2.3 The Graph Convolutional Network

The graph neural network (GNN) architecture is called the graph convolutional network (GCN), where the convolutional neural network (CNN) acts on the graph input . CNN assumes that the data has a natural structure, but the graph data is essentially unstructured. Each node is labeled with some input and output parameters of interest in a weighted and directed graph G. The convolution is performed by $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X$.

The formula above transforms the features in X, asks on a weighted average of the features of its first order neighbors. This transform hasn't applied weight matrix, just aggregating data based on the graph itself. Introducing the weight and bias parameters, it appears a more familiar feed-forward structure:

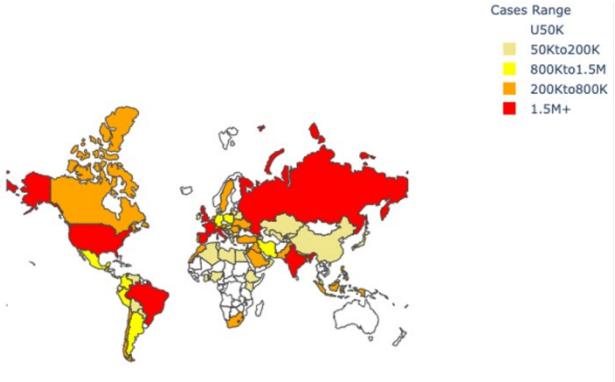
$$A^{(1)} = \sigma\left(Z^{(1)}\right) = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}XW^{(1)} + b^{(1)}\right)$$

Here, σ is just some nonlinear activation. It can continually stack similar layers, until a final output layer. This research only focuses on the graph convolution, thus omit these details. Generally, more convolutions allow each node to aggregate information from higher-order neighbors. Backpropagation is performed similarly to a standard Multilayer Perceptron.

3 EXPERIMENTS

3.1 Description of Data & Preprocessing

The data in this research comes from the statistics of COVID-19 from John Hopkins University and the Figure 1 is the visualization of the Covid 19 Cases in the world [12]. The Johns Center for Systems Science and Engineering (CSSE) released a daily updated version of COVID-19 data on December 22. This dataset contains the number of daily confirmed cases and deaths in every state and county in the United States. This experiment breaks it down into state and county's daily confirmed cases. It is essential to make sufficient data for the training model to ensure accuracy. This study decides to segment actual data. The data from May 1, 2020 to November 30, 2020 is divided into training data sets, and the data from December 1 to December 7 is segmented into test sets. T



The data comes from a detailed map of wearing masks in the U.S from the New York Times. These data contain survey responses of wearing masks, which come from a large number of interviews conducted by the global data and survey firm Dynata at The New York Times' request [13]. These differences reflect disease risk and political differences but may also reflect local characteristics.

To build the model, this research adopts the data of county adjacency from the U.S Census. [14] The county adjacency document lists each county or its counterpart and which counties or counties are adjacent. The

document includes all 50 States, the District of Columbia, Puerto Rico, and the Islands' Territory (American Samoa, The Federation of the Northern Mariana Islands, Guam, and the United States Virgin Islands).

3.2 Modelling the COVID-19 Graph

In infectious disease modeling, multiple time series always represent the observable transmission dynamics values at each location. The prediction dilemma is usually expressed as t – k, t -1, t in a particular time series as a regression learning task and output a single value t + 1 or a future time series t + 1, t + 2, t+3 as a predicted value. According to the spatiotemporal graph model based on motivation [8], this study modified the code from a basic spatiotemporal graph model [15]. This research creates a graph with different edge types for each county in the United States to model the spatial and temporal dependencies. In the spatial domain, edges represent the direct position-to-position movement and are weighted based on the flow rate normalized with the internal flow. Every ordinary node in the spatiotemporal graph has specific characteristics. Each node contains characteristics of past confirmed cases and past deaths as well as the prevalence of masks. This paper takes the number of COVID-19 cases and deaths published by John Hopkins University, including daily reports of new infections and deaths in states and counties in the United States. In addition to the preference of wearing masks, this research utilizes a report provided by the New York Times by many interviews conducted by the global data and survey about wearing masks in each state.

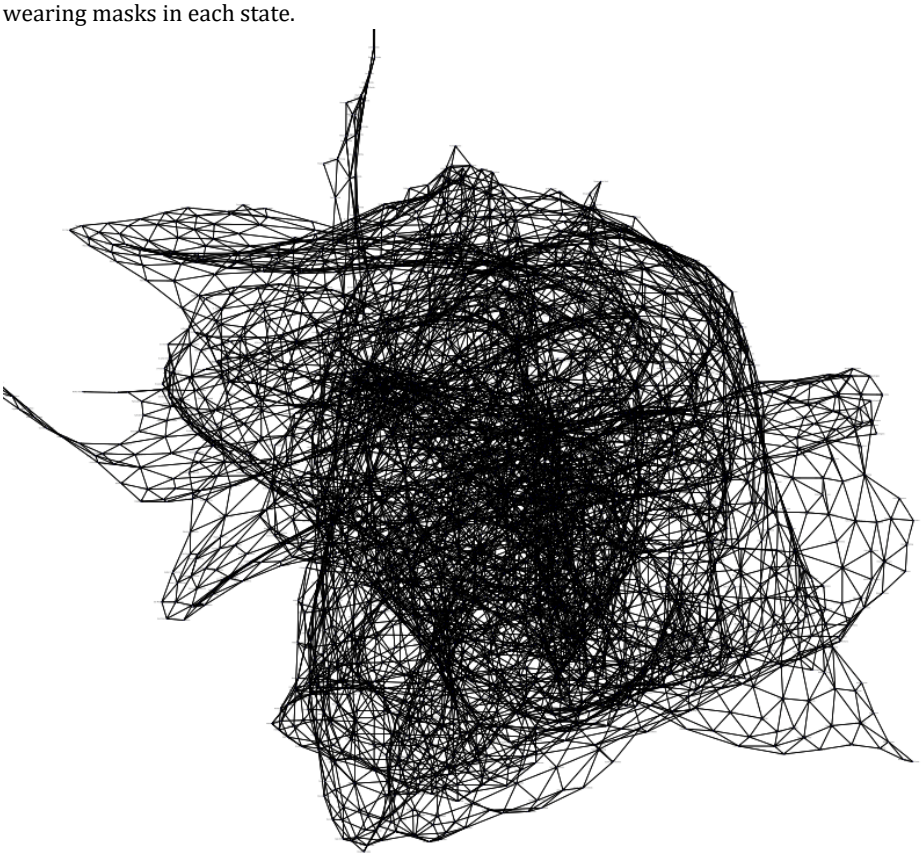


Figure 2: Visualization of the network created by each county from U.S states and their connections

In this model, a graph has many slices, and each node in each slice represents the spatial relationship between counties. And each slice represents a material connection. Figure 2 visualized the network created by counties from states, and each node is represented as the county and connected by their border. Our model has three main layers: the modified GCN layer's initial embedding layer, the standard GCN layer, and the final prediction GCN layer. The initial embed layer performs a graph convolution on the input data and then passes the aggregated data onto the stack. Regularization is performed between each layer using P = 0.5 dropouts. Finally, the prediction layer outputs the data sequence and flattens it into a new node label. The flat data passed to the next layer, and these new node tags pass it to the final prediction layer. The output has two characteristics: confirmed cases and deaths. The loss of the entire network is evaluated using MSE and backpropagation gate error, and Adam is optimized with a learning rate of 1e-3. For backpropagation, it updates the weights after propagated each graph on the network.

4 ANALYSIS ON EXPERIMENTS

This paper adopts the winter solstice data from November 30 to January 22 for testing, and use the winter solstice data from December 6 to November 30 for testing. The goal is to predict a series of dates. The model's output is recursively generated and then feed it back. Figure 3 shows mean squared error of an estimator measures loss of the average squared difference between the estimated values and the actual value after thousand training epochs. As shown below, it can be seen that the model can capture trends the next day, but the data for the next few days do not match the reality. The actual results differ from expectations. Figure 4 is an actual and projected example of the Arizona test period. Also, because the model is not familiar with many situations in the test data, it is sometimes difficult to infer or generalize the model.

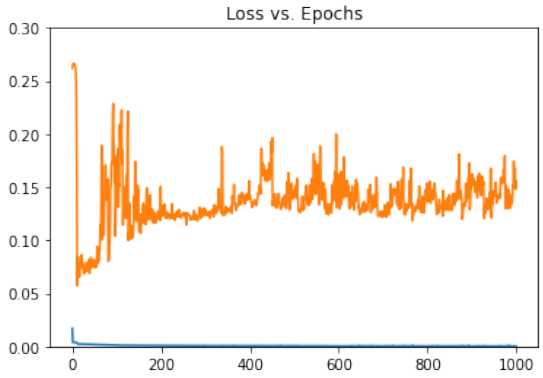


Figure 3: MSE lose by epoch

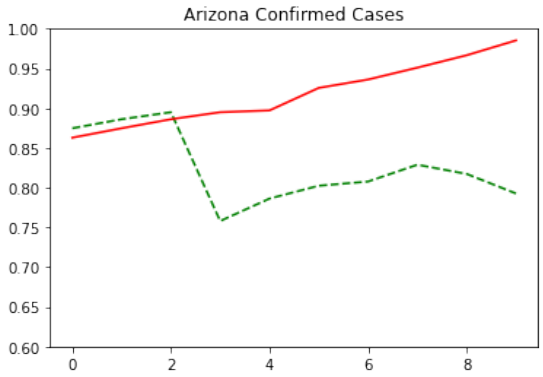


Figure 4: forecasted case in Arizona vs. actual cases

There is still much to be improved, both in COVID-19 and in modeling infectious diseases in general.

The Model has the RMSLE of 0.03627, and an absolute Pearson Correlation of 0.9735. Although the model has shortcomings that did not make reasonable predictions for all days, recent predictions are close to reality. Machine learning requires large data sets. Even this data set contains the data of the whole year, but it still may not be sufficient. And it may improve the model even more by tuning hyper-parameters. These results can be extended by leveraging this new source of powerful mobile information through new techniques in machine learning, incorporating new features, extending the time horizon of long-term predictions, and experimenting with epidemiological flow data from other parts of the world.

5 CONCLUSION

This paper demonstrates the feasibility of adopting a graph neural network to study virus transmission. Although this experiment encounters many problems in the process, it is a powerful tool to help this research on the virus. The researcher can easily extend this modeling framework to regression problems with large-scale spatiotemporal data, disease status reporting, and human movement patterns at a different time and geographic scales in this graph-neural network-based spatiotemporal movement signal based COVID-19 prediction. This model does not rely on assumptions about underlying disease dynamics and can learn from various data, including interregional interactions and region-level characteristics. But this approach is not very useful for studying viral transmission. Some other essential factors in these data sets can inspire different results and functions, such as weather changes. These factors, coupled with increased awareness, can effectively reduce transmission, even if mobility remains constant. Due to various constraints, this experiment can add more features to the nodes to form a comprehensive graph. In subsequent work, it can try to add more pieces of information for the whole composition and expect to take the higher the resolution of the individuals such as hospital, school, community, and the airport as a node. Then given specific proportions to edges, such as the airport and jet flow, the public community's liquidity, and endowed with different proportion and features. If the conditions permit, this experiment can use even traditional viral transmission characteristics to construct this super-high-resolution map with the proportion of the functional situation as the boundary. It will help to understand the model of how the virus spreads and help researchers effectively contain outbreaks before a vaccine can be used on a large scale.

ACKNOWLEDGEMENTS

First, I would like to thank my teachers for providing guidance and feedback throughout this project. Second, I also want to appreciate my partner Yuxin, who had done a bunch of research for collecting references and data.

And my biggest thanks to my tutor Ms. Han, who provided the necessary support, guided me positively, and made me feel confident in my abilities.

REFERENCES

<bib id="bib1"><number>[1]</number>WHO. 2020. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/ (visited on 12/20/2020). (2020).</bib>
<bib id="bib2"><number>[2]</number>Sheryl L Chang, Nathan Harding, Cameron Zachreson, Oliver M Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. arXiv preprint arXiv:2003.10218 (2020).</bib>
<bib id="bib3"><number>[3]</number>Sen Pei and Jeffrey Shaman. 2020. Initial Simulation of SARS. CoV2 Spread and Intervention Effects in the Continental US. medRxiv (2020). https://doi.org/10.1101/2020.03.21.20040303</bib>
<bib id="bib4"><number>[4]</number>James Durbin and Siem Jan Koopman. 2012. Time Series Analysis by State Space Methods: Second Edition (2nd ed.). Oxford University Press.</bib>
<bib id="bib5"><number>[5]</number>CJ Murray et al. 2020. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. (2020).</bib>
<bib id="bib6"><number>[6]</number>Zifeng Yang, Zhiqi Zeng, Ke Wang, SS Wong, W Liang, M Zanin, P Liu, X Cao, Z Gao, Z Mai, et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease 12, 2 (2020).</bib>
<bib id="bib7"><number>[7]</number>Lorch L, Trouleau W, Tsirtsis S, Szanto A, Schölkopf B, Gomez-Rodriguez M (2020) Quantifying the effects of contact tracing, testing, and containment. arXiv preprint arXiv:2004.07641</bib>
<bib id="bib8"><number>[8]</number>Kapoor A, Ben X, Liu L, et al. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. arXiv preprint arXiv:2007.03113 2020</bib>
<bib id="bib9"><number>[9]</number>Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun. Graph Neural Networks: A Review of Methods and Applications arXiv:1812.08434v4 [cs.LG] 10 Jul 2019</bib>
<bib id="bib10"><number>[10]</number>Zonghan Wu, Shirui Pan, Member, IEEE, Fengwen Chen, Guodong Long, Chengqi Zhang, Senior Member, IEEE, Philip S. Yu, Fellow, IEEE. A Comprehensive Survey on Graph Neural Networks</bib>
<bib id="bib11"><number>[11]</number>Diao, Z., Wang, X., Zhang, D., Liu, Y., Xie, K., & He, S. (2019). Dynamic Spatial-Temporal Graph Convolutional Neural Networks for Traffic Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 890-897.</bib>
<bib id="bib12"><number>[12]</number>Anthony Goldbloom, devrishi, Jacky Wang, Peijen Lin, Timo Bozsolik. COVID-19 data from John Hopkins University. https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university. (visited on 12/20/2020). (2020).</bib>
<bib id="bib13"><number>[13]</number>A Detailed Map of Who Is Wearing Masks in the U.S. https://github.com/nytimes/covid-19-data/tree/master/mask-use. (visited on 12/20/2020). (2020).</bib>
<bib id="bib14"><number>[14]</number>United States Census Bureau. County Adjacency File. https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html. (visited on 12/20/2020). (2020).</bib>
<bib id="bib15"><number>[15]</number>THOMAS KIPF. Graph convolutional networks. https://tkipf.github.io/graph-convolutional-networks. (visited on 12/27/2020). (2020).</bib>