The Experience with the REU-sponsored Project on Predicting COVID-19 Pandemics Using Physics-Guided Graph Attention Networks

Yu LIANG*
Department of Computer Science and Engineering, University of Tennessee at Chattanooga
Chattanooga, TN 37403, USA

Dalei WU
Department of Computer Science and Engineering, University of Tennessee at Chattanooga
Chattanooga, TN 37403, USA

*Corresponding Author. Email address: yu-liang@utc.edu

ABSTRACT

The COVID-19 pandemic has significantly impacted most countries in the world. Analyzing COVID-19 data from these countries together is a prominent challenge. Under the sponsorship of NSF REU, this paper describes our experience with a ten-week project that aims to guide a REU scholar to develop a physics-guided graph attention network to predict the global COVID- 19 Pandemics. We mainly presented the preparation, implementation, and dissemination of the addressed project.

The COVID-19 situation in a country could be dramatically different from that of others, which suggests that COVID-19 pandemic data are generated based on different mechanisms, making COVID-19 data in different countries follow different probability distributions. Learning more than one hundred underlying probability distributions for countries in the world from large scale COVID- 19 data is beyond a single machine learning model. To address this challenge, we proposed two team-learning frameworks for predicting the COVID-19 pandemic trends: peer learning and layered ensemble learning framework. This addressed framework assigns an adaptive physics-guided graph attention network (GAT) to each learning agent. All the learning agents are fabricated in a hierarchical architecture, which enables agents to collaborate with each other in peer-to-peer and cross-layer way. This layered architecture shares the burden of large-scale data processing on machine learning models of all units.

Experiments are run to verify the effectiveness of our approaches. The results indicate the proposed ensemble outperforms baseline methods. Besides documented on GitHub, this work has resulted in two journal papers.

Keywords: Layered Ensemble Learning, Physics-guided Learning, COVID-19, Graph Attention Network.

1. INTRODUCTION

During the Summer of 2021, the authors advised an undergraduate on a research project under the sponsorship of the Research Experiences for Undergraduates (REU) program of the National Science Foundation (NSF). This paper is dedicated to presenting the experience with that project entitled "Predict COVID-19 pandemics using physics-guided graph attention networks".

Table I shows the weekly schedule of our project. The REU project team consists of one REU scholar and two advisors. The REU scholar committed the project in full-time from Week 1 through 10.

TABLE I: Weekly schedule of the addressed REU project

Stage I: Project Preparation	
Week -4～-3 Week -2～-1 Week 1 Week 2 Week 3 Week 4	form the team (one REU scholar + two advisors) determine objective and tasks through two-way communication between REU scholar and advisors Training of fundamentals of machine learning Mathematical modeling of pandemic dynamics Literature review of GNN-based pandemics prediction (1) Installation STAN as a baseline GNN-based Covid-19 pandemic prediction system (2) Test STAN using Covid-19 pandemic data as customized dataset.
Stage II: Implementation of the Proposed Project	
Week 5	(1) Data preparation for Covid-19 pandemics status (2) Propose a novel pandemic dynamic equation involving vaccination (3) Incorporate loss function with the proposed pandemic dynamics
Week 6	(1) Formulate graph's edges by international airborne mobility
Week 7 Week 8 Week 9	(2) Introduce attention function in graph network Develop sequential peer-learning framework for GAT Develop layered ensemble learning framework for GAT Introduce attention score into GAT
Stage III: Documentation and dissemination	
Week 10 Week 12	Validate and test the proposed system manuscript preparation for conference paper

The COVID-19 pandemic has rapidly spread to most countries of the world. An accurate prediction/forecast about COVID-19 will enable medical researchers and governments to better respond to this pandemic by informing decisions about pandemic planning, design and distribution of vaccinations, resource allocation, implementation of social distancing measures, and other interventions. Mathematical models (usually described as deterministic or stochastic dynamic systems) and machine learning (ML) models (e.g., deep neural networks, logistic regression, XGBoost, random forests, SVM, and decision trees [1]) form the two dominant categories of prediction techniques.

Mathematical models, which are formulated using partial differential equations (PDE) [2] or ordinary differential

ISBN: 978-1-950492-60-2
ISSN: 2771-5914

https://doi.org/10.54808/IMCIC2022.01.129

equations (ODE) [3] in COVID-19 pandemic simulations, have relatively large error because they are derived based on simplification with high bias. Machine learning models can mirror the randomness of data, but noisy or non-representative data may affect the model's prediction performance and ability of being generalized to unseen data due to high variance. As a result, physics-guided ML models [4] are investigated to optimize the trade-off between bias and variance.

COVID-19 has generated a large volume of data that describes its transmission as a highly connected process. For example, each country's unique dynamic situation can be significantly impacted by other countries' infection trends due to physical interaction between countries and intervention differences (e.g., time of vaccine introduction). Therefore, we need a model that considers the inherent spatial and temporal nature of the disease to incorporate the connections between countries into COVID-19 prediction. Graph neural networks (GNNs) are a particularly useful machine learning method for capturing dependencies between nodes (countries). One type of GNN known as a graph attention network (GAT) further enables nodal interactions to be propagated through a graph structure [5].

Recently, the Spatial-Temporal Attention Network (STAN) [6] model was developed to capture spatial-temporal patterns of COVID-19. The STAN model is a physics-guided GAT that incorporates pandemic physics in the form of the Susceptible-Infectious-Recovered (SIR) model into its loss function to enhance long-term predictions. STAN was tested using COVID- 19 and patient claims data at the county and state levels for the United States, with promising results. STAN was selected as the baseline of our work.

This work extended and optimized STAN at the country level by including more accurate quantification of inter- community interactions and addressing some of the limitations of physics-guided ML models. After improving upon country-level GNN prediction, we addressed the challenge of global COVID-19 prediction. Learning hundreds of underlying patterns from large scale COVID-19 data is beyond a single learning machine, so we present a layered ensemble to predict COVID-19 at the world level. The base layer of the ensemble assigns one GAT (modified STAN model) to each country. The outputs of these GATs are fed into MLPs corresponding to continent and world layers to predict the trend of COVID-19 worldwide.

The reminder of this paper is organized as follows. Section II gives a brief introduction to the preparation of this work. Section III addresses the proposed methodologies of this paper. Section IV presents the experimental results and documentation. Section V concludes the paper.

2. STAGE I: PROJECT PREPARATION

2.1. Training of REU Scholar about the Foundation of machine learning

Data science is inherently interdisciplinary science, which requires the mastery of a variety of skills and concepts, which include Math-foundation (probability/statistics, multivariate calculus, linear algebra, and optimization), computer science (database,), machine learning programming paradigm such as PyTorch, and domain applications (pandemic dynamics). This task took about 10 days because the REU scholar has accumulated solid background in those topics.

2.2. Literature review of physics-guided GNN-enabled pandemic prediction

According to the research topics and the preliminary work of the REU scholar, the literature review mainly focusses on the mathematical modeling, physics-guided machine learning, and graph neural network.

a) ***Mathematical Modeling of Pandemics:*** The base **Susceptible-Infectious-Recovered** (SIR) model from epidemiology [7] is used to model the progression of an epidemic in a closed population over time. The SIR model applied in this work involves vaccinations.

b) ***Physics-Guided Machine Learning:*** Increased popularity of hybrid physics-guided machine learning models ([8, 9, 4]) have been previously implemented to predict COVID-19 trends. We focus on the STAN model developed by [6], a spatio-temporal GAT developed to predict COVID-19 progression based on Johns Hopkins and patient claims data. STAN incorporates the SIR pandemic transmission model into the loss function, and the authors tested their model on USA county and state-level data from 2020, using demographic and geographic data to determine inter-community interactions. However, no known physics-guided ML models have been used to predict COVID-19 progression at the country, continent, and world levels.

c) ***Graph Neural Networks for Pandemic Prediction:*** As investigated by [10], a graph neural network (GNN) is a general framework for a neural network that operates on graph-structured data. Each node has an embedding, or state, with its current feature values, such as a single community at one point in time during a pandemic. The data can be modeled by constructing an attributed graph $G = (V, E)$ where V is the set of nodes representing the set of communities and E is the set of directed edges representing interactions be- tween communities. Defined by [5], graph attention networks (GATs) are a type of GNN that assign learned attention scores to each edge. The attention score between two nodes is a linear transformation weight matrix. The non-linear activation function leaky rectified linear function (LeakyReLU) is applied to compute the attention matrix. Then soft-max function is used to normalize the attention score. Each edge of the node will receive an attention score.

2.3. Design of Technical Contributions

Based on the qualification of the academic background of the REU awardee, the following expected major technical contributions of are determined:

1. Developed a **geographically informed layered ensemble framework** to formulate a multiscale knowledge representation, which ranges from country

through global level, about the COVID-19 trend. As demonstrated in numerical experiments, the ensemble knowledge derived from lower-level networks greatly improves the prediction accuracy of higher-level networks.

2. Developed a **physics-guided Graph Attention Network framework** for country-level COVID-19 predictions, which include (a) incorporating novel pandemic dynamic equations to reduce the limitations of traditional closed- community epidemiological models; (b) introducing mobility-based edge weights, derived from flight counts, to quantify interactions between countries and enhance dependencies between graph nodes; and (c) designing an adaptive physics-guided loss function to optimize the trade-off between ML methods (featured with relatively poor long-term prediction [8, 4]) and mathematical modeling (featured with relatively poor granular prediction [8, 4]).

Using Covid-19 data from Our World in Data [11] as the ground truth data set and flight data from OpenSky Network [12] to model international interactions, the proposed physics-guided graph neural network and layered ensemble frameworks will be critically assessed.

3. STAGE II: PROJECT IMPLEMETATION

3.1. Data Preparation Design

As a computer science senior, the REU scholar was encouraged to a Python program to process and integrate the raw data about Covid-19 pandemic data and airborne mobility data.

The proposed country-level GAT framework is given an input history window of 14 days' worth of feature data. The model predicts the change in the number of active and recovered cases for the next 14 days. Training data spanned 01/2021-03/2021, validation data was 04/2021, and testing data was 05/2021.

COVID-19 data from Our World in Data [11] was processed to create the temporal feature data for the graph nodes. Countries that had no COVID-19 deaths and fewer than 1000 confirmed cases by 01/01/2021 were excluded from the graph. International flight data was collected from Opensky Network [12] to formulate edge weights between nodes.

3.2. SIRVC: Novel SIR Model with Vaccinations and Inter-Community Interaction

Because the REU scholar has Mathematics minor degree, we encouraged her to develop a novel partial differential equation to govern the spread of Covid-19 pandemic disease within and crossing the community.

The canonical SIRV model only models a closed population. To overcome the drawback this work proposed susceptible-infectious-recovered model with vaccinations and inter-community interactions (SIRVC), a novel extension of this model that incorporates dynamic inter-community inter- actions. The SIRVC model assumes that individuals in a community with a disease will interact with individuals outside of that community with the same disease, as is the case in a global pandemic such as COVID-19. The progression of the disease in each sub-community (e.g., country level) is unique due to differences in time of introduction, demographics, vaccination rate, mobility, etc., and affects the progression of the disease in the larger community (e.g., global level).

3.3. Mobility-Based Edge Weights

To formulate the inter-community pandemics spread, we encourage the REU scholar to enrich the edge feature through introducing airborne mobility and attention mechanism [5].

Each edge is directed and represents mobility from one country to another. A directed edge is included from an origin country to a destination country if a flight between the two countries occurred during the time of training (and testing, if known). A directed edge also exists between countries that share a land border to account for ground travel (e.g., the United States and Canada each have a directed edge to the other).

Different from the typical GAT implementation, whose edges are graded by attention scores [5] only (excluding the weight value), our work takes the advantages of both edge weight and an attention mechanism [13]. The edge weights are dynamic and determined by the volume of air travel between two countries.

3.4. Spatial-Temporal Graph Attention Network

STAN model, the baseline framework of addressed project, is already built on spatial-temporal graph attention network. Following the STAN model, we use two GAT layers designed to extract spatial-temporal features from the graph. Historical feature data is concatenated at each time step t and fed into the first GAT layer. Then the modified graph attention mechanism is applied to generate the updated hidden features for each node, which is passed into the second GAT layer. The output of the second GAT layer is passed to the MaxPooling operator to generate an embedding for the whole graph. The graph embedding is input to a gated recurrent unit (GRU) to learn temporal features.

3.5. Adaptive Loss Function

Because the REU scholar has computer science, applied mathematics, and epidemics background, this task is relatively easy.

Machine learning methods are useful for short-term prediction and can learn complex trends from training data. In contrast, mathematical models like SIR are useful for long-term prediction but less accurate for granular day-to-day predictions that involve more interactions than the model can capture. An adaptive physics-guided loss function was proposed to optimize these trade-offs between the short-term and long-term physics violation and data-fitting error. Pandemics physics violation is derived from the modeling and simulation about COVID-19 pandemics.

It should be remarked that we keep a similar neural network architecture as STAN [6] to demonstrate the advantages of our proposed methodology, such as layered ensemble, adaptive physics-guided loss function, etc. We updated the STAN model to include vaccinations and the SIRVC differential equations. However, note that we do not include the rich feature data (patient claims) used by the STAN authors, as the scope of our model is on the global level.

3.6. Two Frameworks for Team Learning

Different from [6], this work proposed a GAT-enabled team learning strategy to get other vertices/countries, which are closely related to the targeting country, involved in the learning process of pandemics pattern. The addressed work proposed two GAT-enabled team learning framework – peer learning and layered ensemble learning.

Compared to the latter, peer learning is much easier to implement. As a result, peer learning framework was developed first so that we can gather enough experience for the layered ensemble learning.

Peer-learning framework: Peer learning indicates all the countries will be trained simultaneously. We have developed a sequential (or cascaded) implementation of GAT- enabled team learning. Future work will focus on the parallel GAT-enabled team learning. In addition, besides the international airborne or ground mobility, some other relations such as geopolitical relation and socioeconomic similarity should also be considered in the construction of peer-learning teams.

Geographically Informed Layered Ensemble Learning: To generate a multi-scale knowledge representation about COVID-19 pandemics, this paper proposed a hierarchical/layered framework to organize the networks by geographic scale. The proposed layered framework trains the graph networks in bottom-up order; the ensemble knowledge derived from the lower-level networks will be fed into the upper-level networks during training and prediction. Such approaches offer advantages like improved data efficiency, reduced over-fitting through shared representations, and fast learning by leveraging auxiliary information [14, 15, 16].

4. STAGE III: DOCUMENTATION AND DISSEMINATION

4.1. Experiment Design

Experiments are designed in the order of increasing difficulty level, which is measured by the data completeness, volume, balance, and integrity.

The original STAN model [6] was compared with baseline models SIR, SEIR, GRU, ColaGNN, CovidGNN, STAN-PC and STAN-Graph, all of which it outperformed. Our modified STAN model, an SIRVC-based adaptive GAT framework for country level predictions, is compared with the STAN model as a baseline. We show country-level results of multiple experiments for the United States and the United Kingdom, two countries with high quality COVID-19 data and high levels of international interactions.

We determined two baselines for our three-layer ensemble to demonstrate the importance of the middle (continent) layer. The first baseline pools all 163 countries to train the world- level MLP to justify the need for the continent layer. The second baseline randomly assigns countries to continents to demonstrate the importance of a geographically informed ensemble.

4.2. Documentation

The aim of our work is to generate a graph-oriented multi-scale knowledge representation, which ranges from county scale through global level, about COVID-19 pandemics using multi-level hierarchical graph networks. The current experiment shows that geographically organizing country-level data can improve the prediction accuracy of higher-level networks.

Besides the knowledge from lower-level network, the cur- rent experiment also demonstrate the benefit of adaptively exploiting the physics knowledge in GAT-based prediction of COVID-19 pandemics.

Furthermore, to improve the country-level GAT framework (and thereby the performance of the higher levels), additional mobility schemes should be incorporated into the graph's edge features. For air mobility, including the number of passengers on each flight would allow for more accurate weights. For ground mobility, including the volume of ground vehicle traffic across international borders would capture more accurate, granular inter-community interactions. To improve the three-layer ensemble, future work should focus on calibrating the country layer with the output of the coarse world-level MLP layer. Additionally, replacing the MLPs of the continent layer with another GAT framework would allow continent-level interactions to be incorporated into predictions at this layer.

4.3. Dissemination

Dissemination was organized in the order of increasing difficulty level. We encouraged the REU scholar to start with a conference presentation and then target at a peer-reviewed journal paper.

We have attempted to publish the results by drafting and submitting manuscripts. As our first attempt, we submitted a manuscript to IEEE Bigdata 2021. By following the submission guidance provided by the conference, the student drafted the manuscript as the first author and then we further polished it as coauthors. After more than one month, we were informed that our paper could not be accepted by the Conference. While all reviewers confirmed that our work was relevant to the conference and of interest to Bigdata users and practitioners, they provided some comments and suggestion on several aspects for potential improvement. For example, they mentioned the weak evaluation due to the use of only two weeks of data and the lack of comparisons to other data- driven models for COVID-19.

After receiving the reviews of IEEE BigData 2021, we improved the original manuscript by incorporating the review comments. Then we submitted the improved manuscript to Neural Processing Letters (NPL). NPL is an international journal that promotes fast exchange of the current state-of-the art contributions among the artificial neural network community of researchers and users. The journal's five-year impact factor is 2.884. Currently our submission is still under review.

5. CONCLUSION AND FUTURE WORK

In this paper, we described our advising and research experience with an NSF REU project on project COVID-19 prediction using physics-guided graph attention networks. During the ten-week project period, we advised an undergraduate who has solid computer science and machine learning background to implement the project. In the research, we devised a country-level GAT framework based upon previous work and proposed a three-layer ensemble, with the GAT framework as the base layer, to predict large-scale COVID-19 trends. The major changes we included to improve the accuracy of existing spatial-temporal GAT frameworks focused on improving the accuracy of internodal interactions using both mathematical formulas and real-world data. After verifying the accuracy of this country-level GAT framework, we trained models for all 163 countries to create the base layer of our ensemble. The second layer used MLPs to represent continents and the third layer used another MLP to perform world-level prediction.

Based on our experimental results, we found that this geographically informed and physics-guided layered architecture can handle large-scale data processing better than a two-layered or non-geographic ensemble. Future work will consist of improving the physics-guided GAT framework and replacing the MLPs of the continent layer with a GAT framework that accurately models continent-level interactions. Additionally, the performance of the entire ensemble may be calibrated by feeding the global predictions back to the country-level layer. We hope that our work demonstrates the power of hybrid and ensemble graph neural networks for pandemic prediction and the utility of these models in informing public health and policy decisions.

In summary, this work is a story-of-success for the training of REU scholar. It will benefit our future REU-oriented projects.

6. ACKNOWLEDGMENT

This work was partially sponsored by NSF REU #1852042. The authors thank Drs. Yingfeng Wang and Mengjun Xie from University of Tennessee at Chattanooga for reviewing this paper. The code was partially developed based on STAN [6].

7. REFERENCES

[1] Nora El-Rashidy et al. “Comprehensive Survey of Using Machine Learning in the COVID-19 Pandemic”. In: Diagnostics 11.7 (2021). ISSN: 2075-4418. DOI: 10. 3390/diagnostics11071155.

[2] Yu Liang et al. “Simulation of the Spread of Epidemic Disease Using Persistent Surveillance Data”. In: proceeding of COMSOL 2010, Boston, USA, October 7-9. Oct. 2010. eprint: https://www.COMSOL.com / paper / simulation- of- the- spread- of- epidemic- disease- using-persistent-surveillance-data-9134.

[3] Deepa Chaturvedi and U. Chakravarty. “Predictive analysis of COVID-19 eradication with vaccination in India, Brazil, and U.S.A”. In: Infection, Genetics and Evolution 92.104834 (2021), pp. 1567–1348. DOI: 10.1016/j.meegid.2021.104834.

[4] Jared Willard et al. “Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems”. In: (2021). arXiv: 2003. 04919 [physics.compph].

[5] Petar Velicˇkovic´ et al. “Graph Attention Networks”. In: (2018). arXiv: 1710.10903 [stat.ML].

[6] Junyi Gao et al. “STAN: spatio-temporal attention network for pandemic prediction using real-world evidence”. In: Journal of the American Medical Informatics Association 28.4 (Jan. 2021), pp. 733–743. DOI: 10.1093/jamia/ocaa322.

[7] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. “A contribution to the mathematical theory of epidemics”. In: Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 115.772 (1927), pp. 700–721. DOI: 10.1098/rspa.1927.0118.

[8] Anuj Karpatne et al. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. 2018. arXiv: 1710.11431 [cs.LG].

[9] Rahul Rai and Chandan K Sahu. “Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques with Cyber-Physical System (CPS) Focus”. In: IEEE Access 8 (2020), pp. 71050–71073. DOI: 10.1109/ACCESS.2020.2987324.

[10] Franco Scarselli et al. “The Graph Neural Network Model”. In: IEEE Transactions on Neural Networks 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008 .2005605.

[11] Edouard Mathieu et al. “A global database of COVID- 19 vaccinations”. In: Nature Human Behavior (May 2021). ISSN: 2397-3374. DOI: 10. 1038 / s41562 - 021 -01122- 8.

[12] Xavier Olive, Martin Strohmeier, and Jannis Lu¨bbe. “Crowdsourced air traffic data from The OpenSky Network 2020”. Version v21.04. In: (May 2021). DOI: 10. 5281/zenodo.4737390.

[13] Liyu Gong and Qiang Cheng. “Adaptive Edge Features Guided Graph Attention Networks”. In: CoRR abs/1809.02709 (2018). arXiv: 1809.02709.

[14] Pedro Avelar et al. “Multitask Learning on Graph Neural Networks: Learning Multiple Graph Centrality Measures with a Unified Network”. In: Sept. 2019, pp. 701–715. ISBN: 978-3-030-30492-8. DOI: 10.1007/978-3-030-30493-5 63.

[15] Sebastian Ruder. “An Overview of Multi-Task Learning in Deep Neural Networks”. In: CoRR abs/1706.05098 (2017). arXiv: 1706.05098. URL: http://arxiv.org/abs/1706.05098.

[16] Yu Zhang and Qiang Yang. “A Survey on Multi-Task Learning”. In: CoRR abs/1707.08114 (2017). arXiv: 1707.08114. URL: http://arxiv.org/abs/1707.08114.