Deep Epidemiological Modeling by Black-box Knowledge Distillation: An Accurate Deep Learning Model for COVID-19

Dongdong Wang,[1] **Shunpu Zhang,** [2] **Liqiang Wang** [1]
[1] Department of Computer Science, University of Central Florida
[2] Department of Statistics and Data Science, University of Central Florida
daniel.wang@knights.ucf.edu, {Shunpu.Zhang, Liqiang.Wang}@ucf.edu

Abstract

An accurate and efficient forecasting system is imperative to the prevention of emerging infectious diseases such as COVID-19 in public health. This system requires accurate transient modeling, lower computation cost, and fewer observation data. To tackle these three challenges, we propose a novel deep learning approach using black-box knowledge distillation for both accurate and efficient transmission dynamics prediction in a practical manner. First, we leverage mixture models to develop an accurate, comprehensive, yet impractical simulation system. Next, we use simulated observation sequences to query the simulation system to retrieve simulated projection sequences as knowledge. Then, with the obtained query data, sequence mixup is proposed to improve query efficiency, increase knowledge diversity, and boost distillation model accuracy. Finally, we train a student deep neural network with the retrieved and mixed observation-projection sequences for practical use. The case study on COVID-19 justifies that our approach accurately projects infections with much lower computation cost when observation data are limited.

Introduction

The spread of infectious diseases is a serious threat to public health and may cause million deaths every year. To effectively battle against infectious diseases, accurate modeling on their transmission patterns is critical. This issue becomes more pressing when the infectious disease, like COVID-19, is unprecedented, transmission dynamics is complex, and observation data are limited. Due to data limitation, we need to solve this problem with the help of conventional physics-based epidemiological models. However, it is still difficult to accurately describe complex dynamics with a single model.

Mixture models are widely used to accurately solve complex transient modeling problems. They can refine temporal scale into several states with different onsets, model these states separately, and then mix modeling results to represent complex dynamics. Although this refinement on temporal scale more accurately depicts the variation in a physical system, the difficulty of calibrating a mixture model and computational complexity can exponentially increase since it can result in very large parameter space, *i.e.*, curse of dimensionality. When prior knowledge about an infectious disease, such as COVID-19, is limited, exhaustive search in such large space is inevitable for accurate model calibration, which can easily render a mixture model impractical. In reality, some modelers propose some assumptions to truncate search space with coarse grid and trade for efficiency and feasibility, but it can cause large uncertainty and model degradation.

To address this problem, we formulate a new approach with black-box knowledge distillation. This approach is developed based on three-fold objectives, including higher prediction accuracy, lower modeling cost, and higher data efficiency. To achieve higher prediction accuracy, we first leverage mixture models to create a comprehensive, accurate, but probably impractical epidemic simulation system. This system is viewed as a black-box teacher model which contains sophisticated modeling knowledge. To reduce modeling cost and make this system feasible, we employ knowledge distillation to transfer the accurate modeling knowledge from this impractical black-box teacher model to a deep neural network for practical use. To realize this knowledge transfer, we collect a set of simulated observation sequences to query the teacher model and acquire their corresponding simulated projection sequences as knowledge. Particularly, for improvement in model performance with limited data, we propose sequence mixup to augment data pool, thus reducing model queries, increasing sequence diversity, and boosting modeling accuracy. With all retrieved and mixed observation-projection sequence pairs, we train a student deep neural network for infection prediction. This student network can perform prediction as accurately as teacher model, but save lots of computation cost, and require fewer observation data.

To the best of our knowledge, we are the first to propose a black-box knowledge distillation based framework to solve epidemiological modeling by leveraging mixture models. Besides this novelty, our work also includes the following contributions: (1) the distilled student deep neural network enables accurate model calibration and projection automatically. (2) Sequence mixup is proposed to reduce teacher model queries for higher efficiency, improve the coverage of obtained data for better accuracy, and further enhance knowledge transfer with fewer observation data. (3) We justify our approach by solving COVID-19 infection projection

and it performs on par or even better than some state-of-the-art methods, like CDC Ensemble, with adequate accuracy over the evaluation period. (4) Our approach provides a general solution to render impractical physics-based models feasible.

Related Work

Epidemiological Modeling

Epidemiological modeling has been extensively studied for decades. It is focused on how to accurately quantify infectious disease transmission dynamics. The proposed methods can be classified into two main categories, classical physics-based modeling and data-driven approach. For physics-based modeling, compartmental modeling, like SEIR (Kermack and McKendrick 1927), is well justified for practical projection. Different from physics-based modeling, thanks to the improvement on data collection, data-driven approaches have been developed based upon statistical modeling on real observation data and widely used for transmission dynamics projection, such as ARIMA(Benjamin, Rigby, and Stasinopoulos 2003) and ARGO(Yang, Santillana, and Kou 2015; Yang et al. 2017). With rapid advances in artificial intelligence, deep learning based modeling as an alternative is proposed to solve infection projection, especially for emergency pandemic like COVID-19 (Wu, Leung, and Leung 2020; Hu et al. 2020; Yang et al. 2020; Fong et al. 2020). However, these data-driven approaches can suffer from observation data limitation. Recently, a hybrid approach named DEFSI (Wang, Chen, and Marathe 2019) adopts compartmental modeling to alleviate data limitation problem in deep neural network training.

Knowledge Distillation

Knowledge distillation (Hinton, Vinyals, and Dean 2015) is widely used to solve deep neural network compression problem. Conventional distillation process is carried out by training a smaller neural network called student model with class probability, which is referred to as "dark knowledge", to retain the performance of original cumbersome ensemble of models called teacher model. This approach can effectively reduce model size, which makes complex models feasible for real-world applications. Many complex applications in computer vision or natural language processing have justified its merits for model size reduction. For example, DistilBERT (Sanh et al. 2019) successfully reduces the size of original BERT model by 40% with maintaining accuracy; TinyBERT (Jiao et al. 2019) leverages knowledge distillation to design a framework for the reduction of transformer-based language model, which leads to the models with lower time and space complexity, thus facilitating its application; relational knowledge distillation (Park et al. 2019) further optimizes distillation process and enables more productive student model, which can even outperform teacher model. However, this effective approach has not been applied to solve complex epidemiological modeling, especially the infeasibility of mixture epidemiological models.

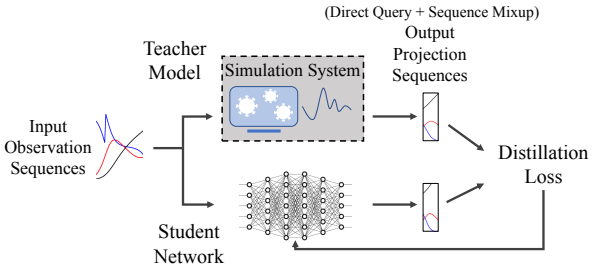

Figure 1: Modeling with black-box knowledge distillation. Teacher model is an accurate but significantly complex comprehensive simulation system. Both observation and projection sequences are simulated results. Model query is optimized by sequence mixup.

Mixup

Mixup is a simple yet effective approach to augment training data and improve model performance (Zhang et al. 2017). This method is proposed to improve the generalization of deep neural network by enhancing coverage of data distribution, especially when training data are limited. The main idea is to incorporate convex combination into data synthesis, which involves mixing features and mixing labels. It has been widely used to address computer vision and natural language processing problems, like Between-Class learning in speech recognition (Tokozume, Ushiku, and Harada 2017) and image classification(Tokozume, Ushiku, and Harada 2018), AutoAugment with learning strategy augmentation for classification (Cubuk et al. 2018), and wordMixup or senMixup with embedding mixup for sentence classification (Guo, Mao, and Zhang 2019). More studies explore its potential for data-efficient learning, such as active mixup (Wang et al. 2020) and ranking distillation in (Laskar and Kannala 2020). However, there is no work using mixup to enhance epidemiological modeling efficacy and efficiency.

Methodology

Figure 1 shows an overview of our approach on epidemiological modeling by black-box knowledge distillation. We leverage mixture models to build a comprehensive simulation system with accurate modeling knowledge yet significantly high complexity. Then, we use simulated observation sequences to query this system to retrieve simulated projection sequences as knowledge. To improve query efficiency and enhance knowledge transfer, sequence mixup is designed to further efficiently augment data pool. With retrieved and mixed observation-projection sequence pairs, a deep neural network is trained to retain the modeling accuracy of the original impractical simulation system and prepared for practical use.

Developing a Teacher Model

Many approaches can be used to create mixture models and build a comprehensive simulation system $\mathcal{M}$. To ensure reliability, we select a widely accepted compartmental model of SEIR as the modeling approach. In SEIR, people in the

modeled society, aka host society, must be in one of the four health states, *i.e.*, susceptible, exposed, infectious, and recovered. The state transition starts from "susceptible", and then moves to "exposed", then to "infectious", and finally reaches "recovered" state. Thus, the model is constrained with the boundary condition of $N = S + E + I + R$, where S, E, I, and R denote susceptible, exposed, infected, and recovered population, respectively, and N represents the population of the entire host society.

For accurate depiction of transient transmission dynamics, we employ linear mixture model (Brauer 2017) to represent the heterogeneity of host society (Bansal, Grenfell, and Meyers 2007). The host society N is divided into several component host communities N_i with the linear combination in Equation 1, and modeling results from these communities will be mixed to represent the dynamics of entire host society N. The division of host society is based on heuristics, which depends on modeling resolution.

$$N = \sum_{i=0}^{n} N_i = \sum_{i=0}^{n} (S_i + E_i + I_i + R_i) \tag{1}$$

Within each community N_i, transmission dynamics can be described by an ordinary differential equation (ODE) system, as shown in Equation 2, across all compartments.

$$\begin{aligned} \frac{dS_i}{dt} &= \alpha N_i - \beta S_i^t I_i^t - \mu N_i S_i^t \\ \frac{dE_i}{dt} &= \beta S_i^t I_i^t - (\sigma + \mu) E_i^t \\ \frac{dI_i}{dt} &= \sigma E_i^t - (\gamma + \mu) I_i^t \\ \frac{dR_i}{dt} &= \gamma I_i^t - \mu R_i^t \end{aligned} \tag{2}$$

where S_i^t, E_i^t, I_i^t, and R_i^t denote susceptible, exposed, infected, and recovered population, respectively, at time t. β, σ, and γ denote infectious, latent, and recovery rate over the entire incidence, respectively. α and μ are referred to as natural birth and death rates during this period, respectively, which are assumed to be zero in this study.

SEIR modeling is a typical boundary value problem (Farlow 1993), the solution of which relies on boundary condition (BC), initial condition (IC), and ODEs. In this study, for each component host community, constant BC is assigned by the total population N_i due to no vital dynamics, IC is determined by the compartment state information $\{S_i^0, E_i^0, I_i^0, R_i^0\}$ at time step $t = 0$, and ODEs are specified by the dynamics coefficients $\{\beta, \sigma, \gamma\}$. Conventional numerical modeling requires model calibration, , which adjusts parameters to obtain agreement between real observation data and modeled results, using grid search for an optimal combination of BC, IC, and ODEs ({BC, IC, ODEs}) within constraints in search space. If the search space for {BC, IC, ODEs} is larger and fine-grained, the calibration results are better fit to the real observation data and simulated projected results are more reliable. Therefore, we construct a comprehensive simulation system with an ensemble of simulation scenarios from large and fine search space, which enables accurate model calibration and projection.

However, the complexity of this simulation ensemble system is very time-consuming for grid search due to curse of dimensionality. For example, suppose we have just 2 options for BC, IC, and ODEs (the real problems require much more). For each component host community, there are 8 simulation scenarios. However, if we have 10 component communities, the ensemble for the entire society N will reach 8^{10} simulation scenarios. It is infeasible to find an optimal solution with random grid search. Therefore, we conduct knowledge distillation to distill this ensemble simulation system into a deep neural network for practical use.

Querying the Teacher Model

Conventional knowledge distillation is carried out by querying the teacher model to obtain prediction probabilities that are referred to as "knowledge". In our problem, the "knowledge" are simulated projection sequences from the simulation system since they contain the features of modeling process. To facilitate acquiring such kind of modeling "knowledge", we conduct model querying as follows. First, we prepare a simulated observation sequence over the calibration period with a {BC, IC, ODEs} for each host community. Each {BC, IC, ODEs} is used as a "key" to query teacher model. Then, the teacher model will use the "key" to return a query answer with a simulated sequence over the calibration and projection period, *i.e.*, a projection sequence. With more queries, more projection sequences are obtained and more accurate modeling knowledge is acquired.

Sequence Mixup

To ensure adequate knowledge, distillation usually requires lots of training data from many model queries. However, too many queries can be time-consuming, and more importantly, the simulated observation sequences are still too limited to acquire diverse knowledge. For improvement in distillation efficacy and data diversity, we employ sequence mixup to reduce the number of queries and enlarge knowledge coverage.

$$\begin{aligned} \hat{x} &= \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n \\ \hat{y} &= \omega_1 y_1 + \omega_2 y_2 + ... + \omega_n y_n \end{aligned} \tag{3}$$

Our sequence mixup is developed with convex combinations of multiple observation sequences x_i and projection sequences y_i with mix rates ω_i, where $\Sigma\omega_i = 1$. Equation 3 presents this mixup process which mixes observation sequences x and projection sequences y in the same manner.

$$\begin{aligned} S^{t+1} &= S^t + \frac{dS^t}{dt} = \sum_{i=1}^{n} \omega_i S_i^t + \frac{d\sum_{i=1}^{n} \omega_i S_i^t}{dt} \\ &= \sum_{i=1}^{n} \omega_i S_i^t + \sum_{i=1}^{n} \frac{d\omega_i S_i^t}{dt} = \sum_{i=1}^{n} \omega_i S_i^t + \frac{d\omega_i S_i^t}{dt} \\ &= \sum_{i=1}^{n} \omega_i S_i^{t+1} \end{aligned} \tag{4}$$

The mixup projection sequence $\hat{y}$ in Equation 3 uses the same coefficients $\omega_1, \omega_2, ..., \omega_n$ as in $\hat{x}$ and it can be briefly

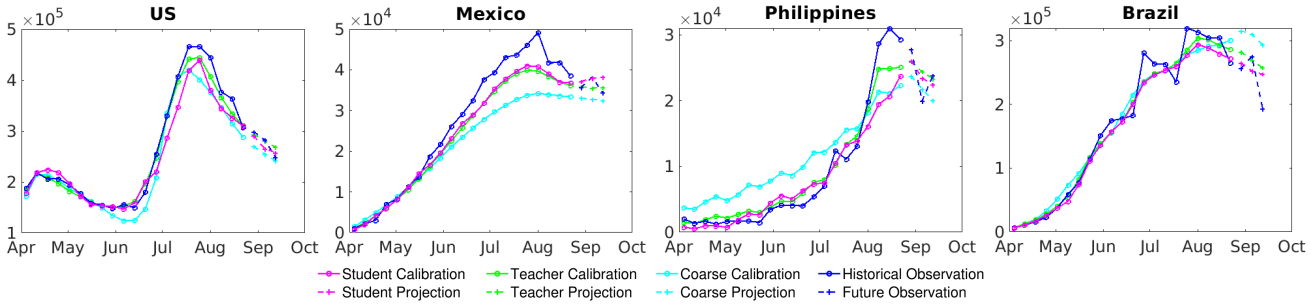

Figure 2: Weekly new infection cases over the calibration (04/06-08/23) and projection (08/24-09/13) periods by teacher model, student network, and coarse search.

proved as follows. Suppose $\hat{x}$ denotes $S^t = \sum_{i=1}^{n} \omega_i S_i^t$ at the current observation time and $\hat{y}$ denotes S^{t+1} at the next projection time. Given the linearity of differentiation, this mixup process $S^{t+1} = \sum_{i=1}^{n} \omega_i S_i^{t+1}$ is justified in Equation 4. Similar proof can be completed for E, I, and R.

These mixed sequences as an alternative to query knowledge efficiently augment training data and enhance the knowledge transfer from teacher model. Thus, all retrieved and mixed sequences construct a training set (X, Y).

Training a Student Deep Neural Network

With the acquired observation-projection sequence pairs (X, Y), a deep neural network is trained to distill the modeling knowledge within the comprehensive simulation system. The conventional distillation process is carried out by minimization on the distillation loss function $L_{dis} = D_1(y_n^{true}, S(x_n)) + D_2(T(x_n), S(x_n))$, where $T(x_n)$ is the output of data x_n from teacher model T, $S(x_n)$ is the output of data x_n from student network S, D_1 is the supervised loss for supervised learning with data label y_n^{true}, and D_2 is the imitation loss for model output imitation. In our problem, there is no knowledge about the true label y_n^{true} for x_n, and thus, the distillation loss is modified to the imitation loss only, as shown in Equation 5. We select mean squared error loss as distillation loss function.

$$L_{dis} = D_2(T(x_n), S(x_n)) \tag{5}$$

The proposed black-box knowledge distillation is a general approach that can be applied to different student networks. In the problem of COVID-19, we use multilayer perceptron (MLP) which is detailed in the case study.

Overall Algorithm

Algorithm 1 presents the overall procedure of our proposed black-box knowledge distillation based epidemiological modeling. Beginning with a modeling approach, a comprehensive epidemic simulation system is built as a teacher model $\mathcal{M}^T$. We then pick a few simulated observation sequences x to query the teacher model and retrieve their simulated projection sequences y. With obtained sequences (x, y), we construct a large observation-projection pool (X, Y) using sequence mixup. Finally, we train a student deep neural network $\mathcal{M}^S$ with (X, Y).

Algorithm 1 Epidemiological Modeling with Black-box Knowledge Distillation

INPUT: A modeling approach F such as mixture SEIR.
INPUT: A set of observation sequences $X_{obs} = \{x_i\}_{i=1}^n$.
INPUT: Hyper-parameters (mixup rate, learning rate etc.)
OUTPUT: A student deep neural network $\mathcal{M}^S$

1: Develop a comprehensive simulation system $\mathcal{M}^T$ based upon F with a set of conditions {BC, IC, ODEs}s
2: With all observation sequences in X_{obs}, query simulation system $\mathcal{M}^T$, retrieve projection sequences $Y_{query} = \{y_i\}_{i=1}^n$, and form an observation-projection pool (X_{obs}, Y_{query}).
3: Construct a mixed sequence pool $(X_{mix}, Y_{mix}) = \{(\hat{x}, \hat{y}) : (\hat{x}, \hat{y}) \in (\sum_{i=1}^n \omega_i x_i, \sum_{i=1}^n \omega_i y_i)\}$ with query results (X_{obs}, Y_{query}), where ω is heuristically chosen.
4: Train a student deep neural network $\mathcal{M}^S$ with $(X, Y) = (X_{obs}, Y_{query}) \cup (X_{mix}, Y_{mix})$ to minimize distillation loss L_{dis}.

COVID-19 Case Study

Experiment Setting

Data. We evaluate our approach on the open COVID-19 datasets provided by Johns Hopkins University (Dong, Du, and Gardner 2020). In this dataset, our experiments are focused on daily infection case increase. With these reported data, we derive active infection cases based on 7-day transmission duration (Thevarajan et al. 2020), as the data do not explicitly report the number of recovered patients. The observation period starts from 04/06/2020 to 08/23/2020 and the evaluation period is from 08/24/2020 to 09/13/2020.

Black-box Teacher Model. A black-box teacher model is built with aforementioned mixture SEIR. The mixture model consists of 10 compartment host communities. Each compartment host community is simulated with 10 choices for N_i to specify constant BC, 2 choices for $\{S_i^0, E_i^0, I_i^0, R_i^0\}$ to specify IC, and 20 choices for each coefficient in $\{\beta, \sigma, \gamma\}$ to specify ODEs. Such choices of parameters are based on heuristics. Most studies on COVID-19 using SEIR model give a wide range of parameter choices (Liu et al. 2020). We refine them to more reliable ranges. With the refined parameter choices, this simulation system contains 160000^{10}

Metric	Model	Calibration				Projection			
		US	Mexico	Philippines	Brazil	US	Mexico	Philippines	Brazil
MAPE	Teacher	**0.0363**	0.1217	**0.3197**	0.0879	**0.0352**	**0.0369**	0.1030	0.1522
	Student	0.0695	**0.1164**	0.3472	**0.0792**	0.0433	0.0527	**0.0984**	**0.1331**
	Coarse	0.0843	0.2269	1.3159	0.1438	0.0727	0.0910	0.1314	0.2923
RMSE (10^5)	Teacher	**0.669**	0.183	**0.101**	**0.790**	**0.209**	**0.028**	0.048	0.703
	Student	1.321	**0.163**	0.163	0.857	0.218	0.041	**0.041**	**0.593**
	Coarse	1.426	0.333	0.229	0.985	0.399	0.063	0.059	1.215

Table 1: Error assessment of model calibration (04/06 - 08/23) and projection (08/24 - 09/13).

scenarios for the entire society N, which is impractical. To facilitate distillation assessment, we conduct random sampling to reduce it to 10^7 scenarios as an approximate version of teacher model to the simulation system for comparative study. The teacher model generates a simulated projection sequence by minimizing the mean squared error between real observation and the simulation over the calibration period, which is similar to exhaustive search.

Query Sequences and Mixup. We randomly pick 1000 {BC, IC, ODEs}s to prepare simulated observation sequences which are used to query teacher system. Note that, compared to the size of the ensemble, this number is so limited that we acquire little knowledge about simulation system with selected sequences, which still follows black-box teacher model setting. Given 1000 query results, we construct a large pool with 100K sequences by sequence mixup, where ω is set heuristically.

Student Deep Neural Network Training. Our student network architecture is an MLP which has 3 hidden layers with 80 neurons each. The batch size is 128 and learning rate is set to 0.1. Adam optimizer is chosen. Weight decay is specified to 1e-5. The total epoch is set to 300 and learning rate is reduced by 90% after every 100 epochs. We select 1K sequences from the constructed sample pool as a training set for efficient training.

Studied Cases. We implement our black-box distillation framework to distill comprehensive infection modeling system for US, Mexico, Philippines, and Brazil. The infection patterns of these countries are representative of complex dynamics which involves multiple peaks and complicates model calibration. To achieve an adequate teacher model on each studied country, we heuristically specify the search space boundaries for {BC, IC, ODEs}s with the information of national population, reported positive cases on March 30th (a week before April 6th), and outbreak severity for each country.

Evaluation Metric. We evaluate infection case modeling performance on both accuracy and efficiency. For accuracy, model calibration and projection are assessed. The performance is quantified by mean absolute percentage error, MAPE $= \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i^o - y_i^m}{y_i^o}|$, and root mean square error, RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^o - y_i^m)^2}$, where y^o is the real observation sequence, y^m is the modeled sequence, and n is the total number of sequences. MAPE and RMSE are two widely adopted metrics to evaluate regression models. While lower MAPE suggests that the general trend is better captured, higher error can occur at larger observation data. RMSE is a better indicator for large values since it offers higher penalty for these errors. Therefore, we use both metrics for accuracy evaluation.

As to computation efficiency, we evaluate model complexity with required simulation scenarios and total time cost for each projection query. For student network, the network training cost is included in each query process although network retraining is not always necessary.

Competing Methods. First, we compare our approach with the approximate teacher model and coarse search to examine accuracy and efficiency. Coarse search is developed upon coarse grid search space for mixture models. We reduce the number of compartment communities to 5, the options for BC to 5, and the choices for each ODE coefficient to 10, which could be taken as a reduced teacher model, but still with the complexity of 10000^5. Similar to teacher model, for practical performance evaluation, we reduce it to 10^5 scenarios with random sampling, which ensures its similar data complexity to student network. In the following sections, approximate teacher model and coarse grid search are referred to as teacher model and coarse search, respectively. Next, we compare our student network with 7 state-of-the-art forecasting models reported from CDC (Bracher et al. 2020). These models are developed with machine learning based methods, like UM and UCLA-SuEIR, statistical methods, like DDS, physics-based model, like JHU-IDD and Columbia, and ensemble approaches, like UVA and CDC Ensemble (Ray et al. 2020).

Results

Accuracy. Our calibration and projection results are reported with weekly increase cases in Figure 2. Student network is comparable to teacher model and significantly outperforms coarse search. These performance differences are quantified with MAPE and RMSE in Table 1. It is shown that, compared to the teacher model, student network achieves similarly low or even lower MAPE and RMSE, over the calibration or projection periods. This observation results from the approximation of teacher model and sequence mixup for student network training. Coarse search yields highest errors due to limited search space.

We compare our student network with 7 state-of-the-art models in Table 2, which are based on the reported data from

Period (from 08/23)	Model							
	CDC Ensemble	UM	DDS	UVA	UCLA	JHU	Columbia	Ours (Student)
1 week ahead	0.0608	0.3866	0.0417	0.0698	0.0367	0.0737	0.0456	**0.0301**
2 week ahead	0.1108	0.0386	**0.0228**	0.0772	0.0889	0.1165	0.0250	0.0623
3 week ahead	0.0581	0.0549	0.0819	0.2724	**0.0077**	0.2572	0.2083	0.0398

Table 2: MAPE comparison of state-of-the-art models and our method on US weekly infection case increase projection between 08/24 and 09/13. The results of other models are collected from CDC, which are reported by COVID-19 Forecast Hub.

CDC (Bracher et al. 2020). Our model consistently outperforms CDC Ensemble, which incorporates all reported state-of-the-art models, with $30\%-50\%$ MAPE reduction over this period. In particular, our model yields more accurate 1 week ahead prediction and more consistent performance over three weeks compared to other models.

Efficiency. From Table 3, student network saves both simulations and time cost by orders of magnitude. Student network and coarse search are on par in total time cost, while the network training takes approximately 300 CPU seconds in our study. This performance gain results from the optimization with sequence mixup and lightweight network design. It justifies that our approach significantly improves modeling efficiency and can facilitate the application of complex and cumbersome epidemiological models.

Significance of Mixup. Sequence mixup, as an efficient method for data augmentation, is very important to enhance knowledge transfer in our approach. Compared to coarse search and teacher model, our student network can learn more scenarios out of search space due to sequence mixup, and this knowledge can overcome the limit from search space, thus even improving calibration and projection accuracy. To justify its importance, we conduct experiments with 100K, 50K, and 25K mixed sequences from 1000 retrieved observation-projection sequences and evaluate their performance difference in calibration and projection for US. From Table 4, the reduction in mixed sequences causes model degradation. The degradation becomes worse in the projection period due to calibration error propagation. Thus, sequence mixup is critical to accurate projection.

Discussion. First, a comprehensive and accurate modeling system is critical in our framework. When this comprehensive teacher model is more complex and accurate, our student network can yield more accurate results. Next, student network can interpolate information in latent space which can resolve space discretization problem in grid search. The space of grid search is often too sparse to find an optimal solution. Therefore, dense search space is imperative,

	Complete Teacher	Approximate Teacher	Student Network	Coarse Search
Simulations	160000^{10}	10^7	10^3	10^5
Time(s)	N/A	$\sim 3\times 10^4$	~ 400	~ 300

Table 3: Model complexity measured by the required simulations and the CPU time cost for one projection query.

	Metric	100K	50K	25K
Calibration	MAPE	**0.0695**	0.0987	0.1459
	RMSE(10^5)	**1.321**	1.831	2.910
Projection	MAPE	**0.0433**	0.1861	0.2813
	RMSE(10^5)	**0.218**	0.985	1.367

Table 4: Calibration and projection errors from student network for US with 100K, 50K, and 25K mixed sequences.

but its cost will exponentially increase. This can be alleviated by our proposed knowledge distillation. In addition, sequence mixup improves training data coverage and boosts model distillation, which helps student network even outperform teacher model. It implies that our proposed knowledge distillation scheme has potential to improve teacher model. Also, if a well-trained student network is obtained, the model could be reused many times, even when new data are included. In contrast, conventional random grid search, like teacher model or coarse search, has to be reset and query all entries again to retrieve projection solutions. This implies student network can save extra query cost.

Conclusion

We propose an innovative accurate modeling approach which leverages mixture models to ensure high accuracy and employs black-box knowledge distillation to reduce complexity and improve accuracy. It consists of teacher model development, model querying, sequence mixup, and student network training. The developed teacher model is a comprehensive simulation system which can accurately model challenged transient dynamics but is impractical. Then, we prepare simulated observation sequences to query this simulation system and retrieve simulated projection sequences as knowledge for distillation. In particular, to save number of queries and enhance knowledge transfer, sequence mixup is designed and effectively augments training data. With retrieved and mixed observation-projection sequences, a student deep neural network is trained as a distilled model for practical use. Our COVID-19 case study on US, Mexico, Philippines, and Brazil justifies that this approach brings in high accuracy but lower complexity. Also, our approach outperforms some state-of-the-art methods, like CDC Ensemble, over the studied period. In future, this work will be extended and applied to more epidemiological studies.

Acknowledgements

This project was support in part by NSF 1704309 and UCF COVID-19 Artificial Intelligence and Big Data Initiative.

References

Bansal, S.; Grenfell, B. T.; and Meyers, L. A. 2007. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface* 4(16): 879–891.

Benjamin, M. A.; Rigby, R. A.; and Stasinopoulos, D. M. 2003. Generalized autoregressive moving average models. *Journal of the American Statistical association* 98(461): 214–223.

Bracher, J.; Ray, E. L.; Gneiting, T.; and Reich, N. G. 2020. Evaluating epidemic forecasts in an interval format. *arXiv preprint arXiv:2005.12881* .

Brauer, F. 2017. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling* 2(2): 113–127.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501* .

Dong, E.; Du, H.; and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20(5): 533–534.

Farlow, S. J. 1993. *Partial differential equations for scientists and engineers*. Courier Corporation.

Fong, S. J.; Li, G.; Dey, N.; Crespo, R. G.; and Herrera-Viedma, E. 2020. Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction. *Applied Soft Computing* 106282.

Guo, H.; Mao, Y.; and Zhang, R. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941* .

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Hu, Z.; Ge, Q.; Jin, L.; and Xiong, M. 2020. Artificial intelligence forecasting of covid-19 in china. *arXiv preprint arXiv:2002.07112* .

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* .

Kermack, W. O.; and McKendrick, A. G. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115(772): 700–721.

Laskar, Z.; and Kannala, J. 2020. Data-Efficient Ranking Distillation for Image Retrieval. *arXiv preprint arXiv:2007.05299* .

Liu, Y.; Gayle, A. A.; Wilder-Smith, A.; and Rocklöv, J. 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine* .

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Ray, E. L.; Wattanachit, N.; Niemi, J.; Kanji, A. H.; House, K.; Cramer, E. Y.; Bracher, J.; Zheng, A.; Yamana, T. K.; Xiong, X.; et al. 2020. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv* .

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* .

Thevarajan, I.; Nguyen, T. H.; Koutsakos, M.; Druce, J.; Caly, L.; van de Sandt, C. E.; Jia, X.; Nicholson, S.; Catton, M.; Cowie, B.; et al. 2020. Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe COVID-19. *Nature medicine* 26(4): 453–455.

Tokozume, Y.; Ushiku, Y.; and Harada, T. 2017. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282* .

Tokozume, Y.; Ushiku, Y.; and Harada, T. 2018. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5486–5494.

Wang, D.; Li, Y.; Wang, L.; and Gong, B. 2020. Neural Networks Are More Productive Teachers Than Human Raters: Active Mixup for Data-Efficient Knowledge Distillation from a Blackbox Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1498–1507.

Wang, L.; Chen, J.; and Marathe, M. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9607–9612.

Wu, J. T.; Leung, K.; and Leung, G. M. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395(10225): 689–697.

Yang, S.; Santillana, M.; Brownstein, J. S.; Gray, J.; Richardson, S.; and Kou, S. 2017. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC infectious diseases* 17(1): 332.

Yang, S.; Santillana, M.; and Kou, S. C. 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences* 112(47): 14473–14478.

Yang, Z.; Zeng, Z.; Wang, K.; Wong, S.-S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; 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(3): 165.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* .