scientific reports

Check for updates

OPEN

COUnty aggRegation mixup AuGmEntation (COURAGE) COVID-19 prediction

Siawpeng Er, Shihao Yang✉ & Tuo Zhao✉

The global spread of COVID-19, the disease caused by the novel coronavirus SARS-CoV-2, has casted a significant threat to mankind. As the COVID-19 situation continues to evolve, predicting localized disease severity is crucial for advanced resource allocation. This paper proposes a method named COURAGE (COUnty aggRegation mixup AuGmEntation) to generate a short-term prediction of 2-week-ahead COVID-19 related deaths for each county in the United States, leveraging modern deep learning techniques. Specifically, our method adopts a self-attention model from Natural Language Processing, known as the transformer model, to capture both short-term and long-term dependencies within the time series while enjoying computational efficiency. Our model solely utilizes publicly available information for COVID-19 related confirmed cases, deaths, community mobility trends and demographic information, and can produce state-level predictions as an aggregation of the corresponding county-level predictions. Our numerical experiments demonstrate that our model achieves the state-of-the-art performance among the publicly available benchmark models.

COVID-19 has been spreading globally and affected almost every country since 2020. In the United States (US), the COVID-19 pandemic started spreading in January 2020, and in March, the daily number of confirmed cases and number of deaths rose to an alarming stage[1]. To control the rapid spread of COVID-19, policymakers in many states imposed movement restrictions and partial confinement to everyone. For example, companies, educational institutes and public places were forced to close or work in remote settings. This has significantly affected the nationwide economy[2] and posed challenges to our daily life. More importantly, the severity of COVID-19 resulted in the loss of lives and might have caused serious long-term complications[3]. With nearly 31 million cases and 0.56 million deaths[4] in the US alone, COVID-19 has become a serious threat to mankind.

Since the beginning of the pandemic, different parties[5–8] have made concerted efforts to collecting and publishing COVID-19 related data, including confirmed cases, deaths, hospitalization information, demographics, and community mobility. Based on the publicly available data, researchers have built predictive models to study the disease dynamics. These efforts include compartmental models such as variants of Susceptible-Infectious-Recovered (SIR) models[9,10], statistical models using regression[11] or time series analysis[12]. Besides, researchers have also applied agent-based simulation modeling[13,14] and deep learning models[15–17] for predicting COVID-19 dynamics. Moreover, the Centers for Disease Control and Prevention (CDC) has been leading a collaborative effort to produce an ensemble model from different research groups[18] (see more detailed discussions of the related work in COVID-19 dynamics prediction in a later section).

Two key measurements used by various research groups in the study of the spread of COVID-19 are the number of confirmed cases and the number of deaths. Both measurements serve to measure the disease dynamics of COVID-19. It should be noted that both the number of confirmed cases and the number of deaths are subject to similar biases that may affect the accuracy of the data, primarily due to reporting criteria[7] or administration delay[19]. However, the number of confirmed cases is often subject to additional measurement error due to the number of people tested and the testing procedure. For example, some datasets[7] only present confirmed cases as the number of people having positive results for a completed polymerase chain reaction test. Such coarse-grained testing numbers may undercount the true spread of the COVID-19, while introducing a significant bias to predictive models. Despite the caveat, the number of deaths is still a relatively better indicator of the intensity of COVID-19 for policymakers to make decisions. As such, we will focus on predicting COVID-19 related deaths in this study.

Furthermore, due to different disease spread severity in different geographical areas, policymakers need to tailor policies based on the particular local situation for the state or the county. Hence, accurate prediction at the

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. ✉email: shihao.yang@isye.gatech.edu; tourzhao@gatech.edu

state and county level is crucial for informed decisions. Indeed, the county-level analysis could provide insightful information at finer granularity for policymakers[16,20–22]. Policymakers can maximize resource allocation efficiency and react promptly in the legislative areas that require urgent attention. Therefore, we aim to build a model which can make short-term predictions for the number of deaths at the county level. We can also produce predictions for the number of deaths at the state level by a simple aggregation of our county-level predictions. Such a model could help both the state and local governments to make an informed decision based on the predictions at the corresponding state and county levels.

The short-term prediction of COVID-19 dynamics is essentially a classical time series modeling problem. For each day, we collect COVID-19 related data as the input, and the desired output would be the predicted number of deaths for the next 2 weeks. On the next day, we obtain the new data from one additional day, and update our predictions for the next 2 weeks starting from the new date. This paper develops a fully data-driven approach to model COVID-19 dynamics. Specifically, we build a self-attention deep learning model, which takes the input data from multiple sources and predicts the number of COVID-19-related deaths for the future 2 weeks at the county level in the United States. The state-level prediction is then obtained by summing up all county-level predictions of the corresponding state. Moreover, we propose a sequence mix-up augmentation approach to further improve the training of our transformer model[23]. We carefully evaluate our proposed methods under different training for different periods. We further compare our best model with other benchmark models to show its strength and usability.

Unlike most existing deep learning approaches based on recurrent neural networks (RNN) such as LSTM-RNN[24], our proposed approach is based on a current state-of-art self-attention model in Natural Language Processing, also known as the transformer model, which is able to capture both short-term and long-term dependencies within the time series and enjoys computational efficiency. In particular, the basic building blocks of the transformer model are the self-attention modules, which directly model dependencies on previous time steps by assigning attention scores. A large score between two events implies a strong dependency, while a small score implies a weak one. In this way, the modules are able to adaptively select time steps that are at any temporal distance from the current time step. Therefore, the transformer model has the ability to capture both short-term and long-term dependencies. Moreover, the non-recurrent structure of the transformer facilitates efficient training of multi-layer models. Practical transformer models can be as deep as dozens of layers, where deeper layers capture higher order dependencies. The ability to capture such dependencies creates models that are more powerful than RNNs, which are often shallow. Also, the transformer models allow full parallelism when calculating dependencies across all time steps, i.e., the computation between any two time steps is independent of each other. This yields a model presenting strong computational efficiency.

Related work. Epidemic prediction is a time series prediction problem. Given a sequence of time with corresponding data, a model needs to predict the target incidence in the future time. There are four main classes of predictive models for epidemic prediction: the compartmental model, simulation modeling, statistical model, and deep learning model. Moreover, for the same class of models, the final prediction may be an ensemble of several other models. In fact, on the CDC website, the final CDC prediction is obtained by ensembling all the submitted models[18].

- *Compartmental model* is one of the most widely used model types for modeling epidemic diseases. It characterizes the disease spread dynamics using systems of ordinary differential equations. One of the most successful compartmental models is the SIR model, which is used to predict disease progression within the population in one area. In the SIR[25,26] model, the population is assigned to Susceptible (S), Infectious (I), or Recovered (R) mode. One variant of SIR model is SEIR model[27–30], which introduces additionally Exposed (E) mode. Other variants of the SIR model include the SIRD[31] model, with additional Deceased (D) mode. Any transitions from one mode to another mode (i.e., the disease spreading dynamics) are modeled as differential equations, often in the form of a transition matrix. Compartmental models are hard to be widely used due to the difficulties to determine the hyperparameters in every differential equation used[32]. One highly successful compartmental model for COVID-19 predictions is from Karlen's group[33]. Karlen's model uses discrete-time difference instead of ordinary differential equations to model the transition matrix.
- *Simulation modeling* uses computer simulation to model different components in the studied environment and observe their interactions. Cellular automata and agent-based simulation are two simulation modeling techniques used to model complex systems[34]. In COVID-19 prediction, several groups[13,14] use agent-based simulation due to its flexibility to simulate dynamic behaviours of systems with large number of entities. While highly flexible, agent-based simulation requires access to extensive computational resources. Besides, multiple simulations are needed for statistically sound observations, resulting in longer inference time. This limitation is noticeable when the simulated systems involve a large number of individual entities.
- *Conventional statistical models* use regression methods to fit the data directly. Such models include ARIMA, Gaussian process regression, and linear regression, which are more flexible than compartmental models. However, most of the time, statistical model usage is limited by the need for more sophisticated hand-crafted features, which often requires knowledge from the domain experts. For example, CLEP model[11] uses an ensemble model of an exponential predictor and a linear predictor. Model from Zhu et al.[21] uses spatio-temporal information among counties to make predictions. Model from Lampos et al.[12] applies statistical model (Autoregressive model and Gaussian Process regression) to predict the disease's dynamics across countries in Europe and the United Kingdom based on online search data.
- *Deep Learning models* are deep neural networks that learn directly from input data. Such models are more flexible than both compartmental models and conventional statistical models. Due to their representation

capability, such models need a less sophisticated handcrafting preprocessing of the input data. In time series prediction problems, some common deep learning models include Long short-term memory (LSTM)[24], Gated Recurrent Unit (GRU)[35], and transformer[36,37]. All the above models can capture intrinsic information from sequential data for accurate prediction. One limitation of such models is the need for large training data. Concurrently with our work, there are other deep learning models including models from Refs.[16,17] that utilize attention mechanism from transformer architecture.

Different models have their merits in their performance for different date ranges. During the beginning of the COVID-19 outbreak, due to the limitations of the available data, we see predictions from the compartmental models or statistical models. With more data available, deep learning models are showing their advantages in the model flexibility and prediction accuracy. It is also more challenging for a model to make accurate predictions as the prediction granularity becomes finer. Intuitively, errors are more likely to be accumulated if a model is tasked to predict more targets than only a few targets. However, finer prediction granularity, such as prediction at the county level, is highly desired since local predictions help policymakers make tailored policies based on individual counties.

Our contribution. In this paper, our goal is to predict the weekly total number of deaths at both county level and state level for the next 2 weeks, given the current week data. Each single-day data include the number of confirmed cases, the number of deaths, community mobility, and population. Our novelty is to connect established Natural Language Processing techniques with the COVID-19 time series prediction. In particular, we build a self-attention model, also known as the transformer model in Natural Language Processing, that is able to capture both the short and long term dependencies within the time series input data. Our model is build upon two main ideas. First, by aggregating the accurate predictions from county level to state level, our model shows strong performance for the state-level prediction task in all prediction periods. Then, we further improve our model, by experimenting with the feasibility of using data augmentation method. We implement mixup[23] as our data augmentation method at the input layer. To the best of our knowledge, this is the first application of mixup data augmentation in COVID-19 data. Data augmentation further improves our model performance, and is particularly helpful when new trends emerge. Using these two core ideas, our proposed method, named COURAGE (COUnty aggRegation mixup AuGmEntation), is able to generate a short-term prediction of 2-week-ahead COVID-19 related deaths for each county and state in the United States. When compared with other benchmark models, COURAGE shows strong performance across different periods, showing its strength and usability for the prediction of the COVID-19 related number of deaths.

Results

To evaluate our proposed method, we compare county-level and state-level predictions using both the county-level and state-level testing sets with mean absolute error (MAE) as our comparison metrics. For county-level predictions, we compare the predictions from our COURAGE model with its two member models—the County model and the Mixup model. We also compare our model with the baseline Naive model, which simply uses the previous week's reported total number of deaths as the prediction. In the state-level predictions, besides the previous mentioned models, we have an additional baseline State model, which is COURAGE prediction generated solely on the state-level input data without any county-level information or the mixup data augmentation. We compare the predictions according to the training period used, corresponding to 0.5, 0.6, 0.7, and 0.8 of the total dataset. We also present the performance of each model across multiple non-overlapping periods. Finally, we compare our models with the available models contributed to the CDC forecast website.

Comparison among models. We summarize our comparison among our models in Table 1. We emphasize that our COURAGE model is an ensemble model that consists of two separate models: the County and the Mixup model. Those two can also serve as standalone models. We have the Naive model (a.k.a. persistence forecast model, which uses the previous week's reported number of deaths as the forecast) as a baseline model in the county-level prediction task. We use both the Naive model and the State model (prediction using only state-level inputs) as baseline models for the state-level prediction task.

The County model and the Mixup model produce more accurate predictions than the Naive model in the county-level prediction task. By combining these two models, we improve our COURAGE model's prediction accuracy. We also get accurate predictions from the County model and the Mixup model in the state-level prediction task. They outperform both the Naive model and the State model. When combining both models, COURAGE has the best accuracy under different prediction periods, shy from the Week 2 predictions when using the training dataset dated from March 7, 2020, to December 1, 2020. We observe that the County model and the Mixup model have their strengths for different periods. Our COURAGE model often obtains the best of both models and produces the most accurate predictions.

Predictions for different periods. We show the number of deaths prediction for different periods in Table 2. Since we use different amount of datapoints as our training set, our training data consists of information across different periods. We take the non-overlapping period from each testing set as a separate out of the sample prediction period. In all the periods, the County and Mixup models produce better predictions than the Naive model. The result shows the feasibility of both models. For the Mixup model, its strength is visible in the last two periods of prediction. In the last prediction period, we use our trained model (using training set with data from 2020-03-07 to 2020-12-01) to predict the latest data with a prediction period from 2021-01-18 to 2021-03-14. Our model is able to predict accurately for the new data, as illustrated by plot of prediction for New York in Fig. 1

Training size	Method	County level		State level	
		Week 1	Week 2	Week 1	Week 2
0.5 (2020-03-07 to 2020-08-22)	Naive	2.0000	2.2765	51.5932	71.8391
	State	–	–	64.1419	81.9421
	County	1.9205	2.2808	48.0625	67.1583
	Mixup	1.9227	2.2442	51.4493	63.4738
	COURAGE	1.8691	2.1602	47.7390	60.8701
0.6 (2020-03-07 to 2020-09-24)	Naive	2.2302	2.5698	58.7599	84.1373
	State	–	–	64.1419	81.9421
	County	2.1805	2.5331	53.0127	73.8226
	Mixup	2.0932	2.3526	52.5268	70.6528
	COURAGE	2.0984	2.3633	51.1680	68.7564
0.7 (2020-03-07 to 2020-10-28)	Naive	2.6275	3.0295	71.8298	102.1820
	State	–	–	101.2264	123.0914
	County	2.4912	2.8654	65.3611	85.1719
	Mixup	2.5054	2.8034	66.9685	83.6469
	COURAGE	2.4583	2.7547	64.9884	81.9787
0.8 (2020-03-07 to 2020-12-01)	Naive	3.0036	3.3499	80.5957	106.7175
	State	–	–	118.9543	147.5468
	County	2.8240	3.1477	74.7763	98.2125
	Mixup	2.8258	3.0506	71.6854	84.4188
	COURAGE	2.7670	2.9753	70.8481	86.3737

Table 1. Prediction of county-level and state-level weekly total number of deaths. The comparison metrics used is MAE, and the testing dataset starts from the 2020-08-23, 2020-09-25, 2020-10-29 and 2020-12-02 to 2021-02-07 for each corresponding training dataset.

Prediction period	Model	Week 1	Week 2
2020-08-23 to 2020-09-24	Naive	26.6183	28.9819
	County	23.2328	25.2783
	Mixup	25.3481	26.5330
	COURAGE	23.5522	25.3955
2020-09-25 to 2020-10-28	Naive	27.6227	41.1483
	County	24.0228	31.3968
	Mixup	27.0248	37.3008
	COURAGE	24.1806	32.2430
2020-10-29 to 2020-12-01	Naive	59.7121	95.9124
	County	48.3641	62.4764
	Mixup	51.0948	64.5994
	COURAGE	49.1359	61.9947
2020-12-02 to 2021-01-17	Naive	80.5957	106.7175
	County	74.7763	98.2125
	Mixup	71.6854	84.4188
	COURAGE	70.8481	86.3737
2021-01-18 to 2021-03-14 (use last trained model)	Naive	84.5587	122.9295
	County	75.8700	82.0913
	Mixup	76.3645	80.7036
	COURAGE	75.0621	79.7467

Table 2. Different prediction periods for the weekly total number of deaths. The prediction metrics reported is MAE.

and additional plots for other major states (Illinois, California, Texas, Arizona) in Supplementary Information. We mark in Fig. 1 with light grey lines for different prediction date ranges as that of Table 2. In the first two periods, the number of deaths is relatively low and stable. There is an increase in the number of deaths in the third period. Finally, we can see a huge increase or decrease in the number of deaths in the last two prediction periods. In scenarios with relatively stable trends, the County model provides better predictions than that of the Naive

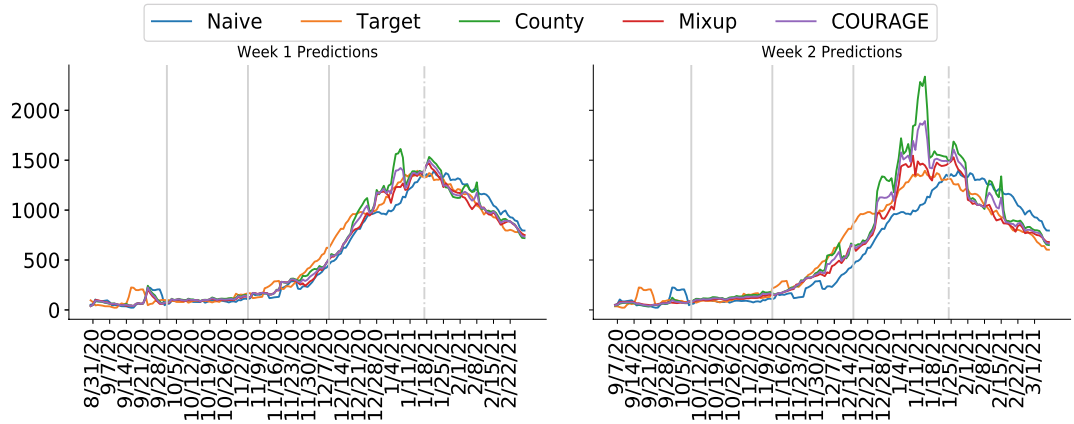

Figure 1. New York's weekly total number of deaths for Week 1 (left) predictions and Week 2 (right) predictions. Vertical lines separate different prediction periods as in Table 2. The last dashed vertical line marks the prediction period of recent data using our last trained model. "Target" is the true reported number of deaths of New York. More plots for other major states are presented in Supplementary Information.

Model	Week 1	Week 2	Average
Karlen[33]	56.2163	77.3440	66.7801
Ensemble[18]	57.3227	79.1330	68.2278
COURAGE	61.1206	76.9096	69.0151
Mixup	61.7890	77.1330	69.4610
UMass-MB[43]	57.1578	84.5514	70.8546
Oliver Wyman[40]	58.9645	85.6277	72.2961
MOBS[45]	60.4486	84.9539	72.7012
County	62.7057	83.5585	73.1321
GT-DeepCOVID[15]	67.1962	90.1549	78.6756
USC[46]	68.1082	93.3954	80.7518

Table 3. Comparison among different models for average MAE (from 2020-11-07 to 2021-02-06).

model. With a more drastic change in the trend, a model trained with mixup data augmentation generalizes better, hence outperforming both Naive and County models. When combining predictions from the County and Mixup models, COURAGE model provides a balanced and more accurate prediction across different periods.

Comparison with benchmark models. We obtain all publicly available predictions from the CDC website for comparison[9,15,18,33,38–51]. Different research groups or individuals contribute towards these predictions. CDC compiles predictions from every model that submits their weekly predictions and uploads them to the CDC website. Note that all prediction models have their strengths in different date ranges. Moreover, they outperform most of the simple methods. We compare our models' predictions with those who send predictions to the CDC official website. The comparison is done for predictions from November 11, 2020 to February 6, 2021. For each model, we calculate MAE for Week 1 and Week 2 predictions, and we calculate the average MAE. We summarize the result in Table 3. Then we rank each model using the average MAE obtained for both Week 1 predictions and Week 2 predictions. Due to the extensively long list of models, we only show results from the top 10 models in both tables. From Table 3, we can observe that our County model produces competitive accuracy for the number of deaths prediction. Our County model is a top 10 model in terms of prediction accuracy. When we augment our dataset, our Mixup model further improves the prediction accuracy. We wish to examine which period mixup data augmentation contributes the most to model accuracy. Mixup data augmentation improves our model when a new trend emerges and helps our model to achieve better prediction in the last period, as illustrated in Table S1 of Supplementary Information. By combining strengths from the County and Mixup models, our COURAGE provides a good balance across different prediction periods.

Discussion

From our results, we see that our COURAGE model gains its strengths from both member models, which are the County model and the Mixup model. The two core ideas associated with these member models are the aggregation of the high-quality county predictions (County model) and data augmentation (Mixup model).

The aggregation of high-quality county predictions results in high-quality state predictions. We could see high-quality county-level predictions from County and Mixup methods from Table 1. This result shows our model and training are effective in extracting and utilizing the information from the input data. When we aggregate these accurate county-level predictions, our models can produce state-level predictions that outperform baseline models. As an illustrative comparison, the State model (baseline model) uses only state-level data as input. Such input is limited and of coarser grain, making it harder for the State model to produce accurate predictions. As a result, the State model's performance pales when compared with our model, as well as the Naive model. One way to refine state-level prediction is through the use of county-level data. By training our models using the larger county-level dataset, we obtain accurate county-level predictions. When we sum these high-quality county-level predictions, we obtain accurate state-level predictions. We justify our intuition from state-level predictions of Table 1, by showing a strong result from our model.

The data augmentation from Mixup method also played a significant role in improving the accuracy of our predictions. Due to the fast-evolving dynamics of COVID-19, most models that submit their predictions to the CDC have high accuracy only for a certain period. Existing trends are much easier to fit than emerging trends, and any changes of an existing trend are hard to predict due to the scarcity of data available at the onset of changes. The fundamental reason for prediction deterioration is the lack of visibility of the new trend given the limited data available. If we could increase the amount of available data by incorporating new data, predictive models could be improved when a new trend emerges. However, data generation is hard for any new trend. One way to improve the model's prediction on emerging trends is to improve its generalization with augmented input data. Our Mixup model uses a well-proven data augmentation technique to create new input data. Such training helps our model achieve accurate predictions in the emergence of unseen data.

By combining the strength of each member model (the County model and the Mixup model), we obtain the COURAGE model. Our COURAGE model obtains its prediction by averaging County prediction and Mixup prediction. This simple ensemble method allows our COURAGE model to decrease the variance in different models and achieve a balanced model.

While our COURAGE model shows strong results, one limitation of our current model is that the self-attention matrix from the encoder is not easily translated to an explainable pattern. Our model's predictions use information from the confirmed cases, deaths, population, and mobility data. Their interaction is encoded by the encoder, with self attention as a key mechanism. Hence, it will be beneficial to the research community if we could see any important relationship among these inputs through the attention matrix. Besides, our model is a relatively concise model that leverages temporal data. In our future work, we plan to extend our work to include spatial information such as interaction among counties or major cities. We expect that the inclusion of such geographical information would further improving our model's predictions. Currently, our model predicts 2 weeks ahead predictions. We plan to include predictions of a longer horizon up to 4 weeks ahead predictions, and also generate a probabilistic forecast that explicitly accounts for forecast confidence.

Methods

In this section, we present our data sources and data processing used in this paper. We also present details of our transformer-based model, mixup data augmentation technique, and training procedure.

Data sources. We use three comprehensive datasets in this study, including confirmed cases, deaths, population, and community mobility from two sources. We only focus on states in the mainland of the United States and do not consider Hawaii, Alaska, and other unincorporated territories in this paper. We use data from 47 states and 3206 counties.

Confirmed cases and deaths of Covid-19 We use data from the JHU CSSE Covid-19 dataset[6]. This publicly available data is a curated dataset from different sources. The data used is collected from January 22, 2020, to February 7, 2021. We use confirmed cases and deaths from the dataset for every county from 47 targeted states.

Community mobility Mobility data have been shown to help the risk analysis of COVID-19 and enhance the predictions of the COVID-19 related deaths and cases[21,52–56]. In current article, we use Google mobility data[8] to incorporate mobility information in our model. These reports record a community's daily movement by county for different areas such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. There is a baseline day's value for every day of a week, which represents the usual community mobility value. The baseline day's value is the median of the data from January 3, 2020, to February 6, 2020. Google mobility data reports the movement pattern of the US community. Each day's data is in the percentage of changes from the baseline day's value. A negative value for a particular day indicates a decrease in time spent at the area from the baseline day, while a positive value indicates an increase in mobility from the baseline day.

Demographic information We use population information from the JHU CSSE Covid-19 dataset[6] which includes population data for each corresponding county.

Data preparation. We retrieve features required for our model training, including the number of confirmed cases, the number of deaths, population, and mobility information from our datasets. We consolidate all input data into state-level and county-level datasets. We also include smoothed (average over past 7 days) confirmed cases and deaths as our input features. We separate total data into training and testing datasets for each corresponding level. To test our method for a different period, we try different amounts (50%, 60% and 70%, and 80%) of the total data as our training dataset, leaving the remaining data as our testing dataset. Since we use multiple features as inputs, we apply standardization to the inputs to accommodate differences in scale for each input. We also use additional data from February 8, 2021, to March 14, 2021, to test the model trained

using 80% of the data in order to examine whether our trained model can continue to perform well for a new period without additional training.

Transformer-based model. The prediction of COVID-19 related number of deaths given a sequence of input is a time series modeling problem. In a typical time series prediction setting, given a sequence of previous days' number of deaths, the aim is to predict the number of deaths for the next future day. For our current article, our prediction problem is different from this typical time series problem. In the current article, our input consists of the current week's number of deaths, number of confirmed cases, smoothed (average over 7 days) number of confirmed cases, smoothed number of deaths, community mobility data, and population of the area. More importantly, instead of predicting the daily number of deaths, our model predicts the weekly total number of deaths for the next 2 weeks (Week 1 and Week 2), given the current week (Week 0) input data. We join the current week input information together to form a data vector of dimension K. The input for our COVID-19 prediction problem can be viewed as a sequence of K-dimensional vectors. Suppose we are given a sequence $\mathcal{S} = \{k_j\}_{j=1}^{L}$ of L days data, where each single-day data $k_j \in \mathbf{R}^K$, occurs at time j.

The key ingredient of our transformer-based model is the self-attention module. Different from RNNs, the attention mechanism does not have recurrent structure. In order to incorporate the temporal information into the inputs, we use the original positional encoding method[36] to our data vector. Alternatively, we could also use other positional encoding methods such as relative position[57] to provide the temporal information for each of single-day data vector in our input sequence.

Before passing to the attention module, we first transform our sequence of single-day data vectors using a matrix $\mathbf{U} \in \mathbf{R}^{M\times K}$. After the transformation, for any single-day data x_j and its corresponding time stamp j, the temporal vector z_j and the single-day data vector $\mathbf{U}k_j$ both reside in $\mathbf{R}^M$. Given a sequence of L days data $\mathcal{S} = \{k_j\}_{j=1}^{L}$, we get

$$\mathbf{X} = (\mathbf{UE} + \mathbf{Z})^T, \tag{1}$$

where $\mathbf{E} = [k_1, k_2, \ldots, k_L] \in \mathbf{R}^{K\times L}$ is the sequence of single-day data vectors, $\mathbf{Z} = [z_1, z_2, \ldots, z_j] \in \mathbf{M} \times \mathbf{L}$ is the concatenation of the temporal vectors.

We pass $\mathbf{X}$ through the self-attention module. Specifically, we compute the attention output $\mathbf{S}$ by

$$\mathbf{S} = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{M_K}}\right)\mathbf{V}, \text{ where } \mathbf{Q} = \mathbf{XW}^Q, \mathbf{K} = \mathbf{XW}^K, \mathbf{V} = \mathbf{XW}^V. \tag{2}$$

Here, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are the query, key and value matrices obtained by different linear transformations of $\mathbf{X}$, and $\mathbf{W}^Q, \mathbf{W}^K \in \mathbf{R}^{M\times M_K}, \mathbf{W}^V \in \mathbf{R}^{M\times M_V}$ are weights for the respective linear transformations. In practice, multi-head self-attention increases model flexibility and is beneficial for data fitting. In multi-head self-attention, different sets of weights $\{\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V\}_{h=1}^{H}$ are used to compute different attention outputs $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_H$. The final attention output for the sequence is then obtained by concatenating all the attention outputs and passing through the final linear transformation.

$$\mathbf{S} = [\mathbf{S_1}, \mathbf{S_2}, \ldots, \mathbf{S_H}]\mathbf{W}^O \tag{3}$$

where $\mathbf{W}^O \in \mathbf{R}^{HM_V\times M}$ is an aggregation matrix.

We highlight that the self-attention mechanism allows the selection of any single-day data whose occurrence time is at any distance from the current time. The j-th column of the attention score from the Softmax $(\mathbf{QK}^T/\sqrt{M_K})$ indicates the extent of dependency of j-th single-day data (k_j) on its history. Hence, attention mechanism allows the capturing of short and long term dependencies of the sequence data. On the other hand, RNN-based models encode the data's history sequentially via hidden representations of events, where the state of j depends on that of $j-1$, which in turn depends on $j-2$, etc. If the RNN fails to learn sufficient information for single-day data at j, subsequent hidden representation of any other single-day data at t where $t \geq j$ will be adversely impacted.

The attention output $\mathbf{S}$ is then fed through a position-wise feed forward neural network, generating a hidden representation $\mathbf{h}(j)$ of the input data sequence:

$$\mathbf{H} = \text{ReLU}(\mathbf{SW}_\mathbf{1}^\mathbf{FF} + \mathbf{b}_1)\mathbf{W}_\mathbf{2}^\mathbf{FF} + \mathbf{b}_2, \ \mathbf{h}(j) = \mathbf{H}(j, :). \tag{4}$$

Here $\mathbf{SW}_\mathbf{1}^\mathbf{FF} \in \mathbf{R}^{M\times M_H}$, $\mathbf{SW}_\mathbf{2}^\mathbf{FF} \in \mathbf{R}^{M_H\times M}$, $\mathbf{b}_1 \in \mathbf{R}^{M_H}$,$\mathbf{b}_2 \in \mathbf{R}^M$ are the corresponding weights and biases of the feed forward neural networks. The resulting matrix $\mathbf{H} \in \mathbf{R}^{L\times M}$ contains hidden representations of all the information in the input sequence, where each row corresponds to a particular information. We use this final representation as an input to our linear decoder layer and obtain our predictions of the weekly total number of deaths for next 2 weeks.

In a typical time series prediction setting, the number of deaths prediction only forcasts the next day given the current week data. In such typical time series prediction, we need to implement masking for the attention mechanism to prevent "peeking into the future" issue. The masking allows any j-th data to attend only to any t-th data where $t \leq j$. In the current article, our model is predicting the weekly total number of deaths for the next 2 weeks (Week 1 and Week 2), given the current week (Week 0) input data. This setting frees us from such masking requirement since the model is implicitly masked from accessing the future total number of deaths from the current week data.

A transformer based model allows us to stack multiple self-attention modules together, and inputs are passed through each of these modules sequentially. In this way our model is able to capture high level dependencies.

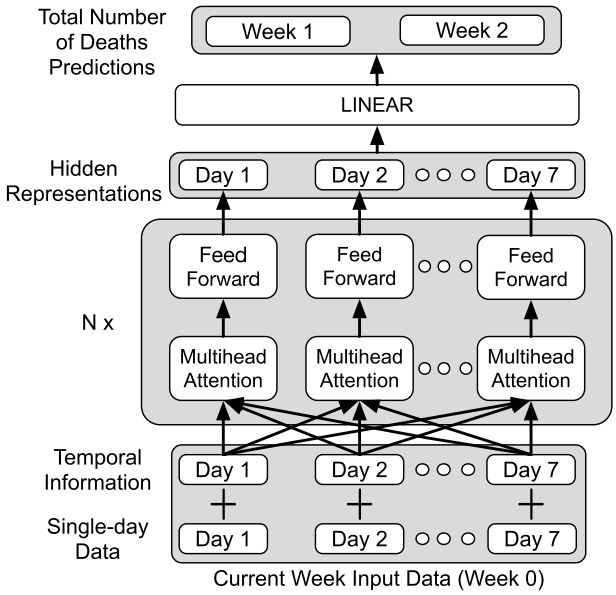

Figure 2. Transformer based model architecture.

We remark that stacking RNN/LSTM are susceptible to gradient explosion and gradient vanishing, rendering the stacked model more difficult to train. Figure 2 illustrates the architecture of our transformer-based model used in this project.

Mixup data augmentation. Data augmentation is a commonly used technique to improve the deep learning models' generalization. One recent augmentation method is mixup[23,58,59] data augmentation. In mixup data augmentation, given $\mathbf{X}$ as the input space of total training data and $\mathbf{Y}$ as the corresponding output values space, each training set is a pair of (x_i, y_i). Mixup data augmentation constructs new data by interpolating it from existing data.

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

where (x_i, y_i) and (x_j, y_j) are two examples drawn at random and $\lambda \in [0, 1]$. In our model, we use mixup data augmentation at input layer.

Training objective. We train the transformer model using the Huber loss function. Specifically, the training objective is defined as

$$\min_{\theta} \ell\big(f_\theta(\mathbf{X}), \mathbf{Y}\big) = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \text{where } z_i = \begin{cases} \frac{1}{2}(f_\theta(x_i) - y_i)^2, & \text{if } |f_\theta(x_i) - y_i| \le \delta \\ \delta\left(|f_\theta(x_i) - y_i| - \frac{1}{2}\delta\right), & \text{otherwise,} \end{cases}$$

$\mathbf{X}$ and $\mathbf{Y}$ are the input space and the target space, with a pair of testing sample as (x_i, y_i). We have n samples, and δ is a tuning hyperparameter.

Training details. We use the transformer-based model (refer to "Transformer-based model" section) for predicting both county-level and state-level number of deaths. Specifically, we use a transformer encoder of 32 model dimensions, 1 encoder layer with 8 attention heads and 64 feedforward dimensions. The output of the encoder layer connects to a single linear layer decoder for predicting both weekly total number of deaths for next 2 weeks (Week 1 and Week 2) using the current week input data. We use Adam[60] optimizer for its superior empirical performance in training a neural network and set 0.001 as our initial learning rate. For every 100 epochs, we decay the learning rate by half for a total of 500 epochs of training. Figure 3 illustrates our training process.

We use both the state-level and county-level training sets to train our models. We use the smoothed number of deaths as our prediction target during training. Upon completion, all the models are used to predict weekly total number of deaths for the next 2 weeks (Week 1 and Week 2) using county-level dataset. We sum all the county predictions of the corresponding state as the predictions of that state. We denote the model trained as the County model. For the Mixup model, there is an additional mixup data augmentation applied to the input layer during the training phase. Our COURAGE model is an ensemble model of two member models, the County

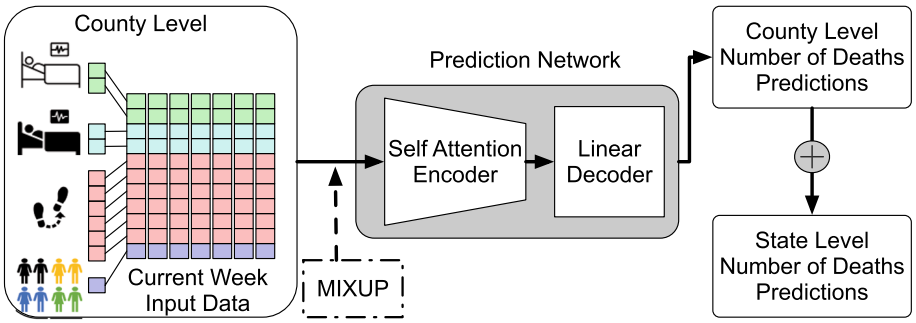

Figure 3. Overview of prediction flow. The county-level and state-level predictions for weekly total number of deaths are for the next week (Week 1) and the second week (Week 2) from the current week (Week 0) input data.

model and the Mixup model. COURAGE takes the average of both predictions from the County model and the Mixup model as its prediction. Our baseline model is a Naive model that takes the current week total number of deaths as both Week 1 and Week 2 total number of deaths predictions. For the state-level prediction task, we have another baseline model—a State model that uses only state-level data as input when making the predictions.

Conclusion

In summary, this article presents the new model COURAGE for COVID-19 predictions at county level and state level for the United States. We use county-level data to train COURAGE and obtain state-level predictions through aggregating high quality county-level predictions. We improve our model using mixup augmentation and ensemble predictions from both the County and Mixup models as our final output. To the best of our knowledge, our model is the first model that use mixup data augmentation to improve the accuracy of COVID-19 related number of deaths prediction. Our experiment shows that this new application of mixup data augmentation helps improve the model's prediction accuracy when new trends occur. COURAGE is a flexible model, that each member model (the County model and the Mixup model) can be used as a standalone model to produce accurate predictions in different periods. When both member models are ensembled together, COURAGE achieves accurate predictions across all periods. COVID-19 is a serious crisis affecting our daily life and economy. Accurate predictions of disease dynamics is a challenging task, especially when new trends emerge. We hope that through our new training method, we can improve COVID-19 number of deaths predictions and provide insight for resource allocation and disease control planning.

Received: 3 May 2021; Accepted: 25 June 2021
Published online: 12 July 2021

References

1. CDC data tracking. https://covid.cdc.gov/covid-data-tracker.
2. COVID-19 Economic Crisis. https://carsey.unh.edu/COVID-19-Economic-Impact-By-State.
3. Long-Term Effects of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html.
4. Coronavirus in U.S.:Latest Map and Case Count. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html (Accessed 7 Apr 2021).
5. Times, The New York. Coronavirus (Covid-19) Data in the United States (2021).
6. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
7. COVID Tracking Project. https://covidtracking.com/.
8. Google LLC. Google COVID-19 Community Mobility Reports. https://www.google.com/covid19/mobility/ (Accessed 16 Mar 2021).
9. COVID-19 Simulator. https://covid19sim.org/documents/policy-methods/.
10. Interpretable sequence learning for COVID-19 forecasting. https://cloud.google.com/solutions/interpretable-sequence-learning-for-covid-19-forecasting.
11. Altieri, N. *et al.* Curating a COVID-19 data repository and forecasting county-level death counts in the United States. *Harvard Data Sci. Rev.* https://doi.org/10.1162/99608f92.1d4e0dae (2020).
12. Lampos, V. *et al.* Tracking COVID-19 using online search. *npj Digit. Med.* https://doi.org/10.1038/s41746-021-00384-w (2021).
13. Kerr, C. C. *et al.* Covasim: An agent-based model of COVID-19 dynamics and interventions. *medRxiv* https://doi.org/10.1101/2020.05.10.20097469 (2020).
14. Germann, T. C. *et al.* Using an agent-based model to assess K-12 school reopenings under different COVID-19 spread scenarios—United States, school year 2020/21. *medRxiv* https://doi.org/10.1101/2020.10.09.20208876 (2020).
15. Rodríguez, A. *et al.* DeepCOVID: An operational deep learning-driven framework for explainable real-time COVID-19 forecasting. *medRxiv* https://doi.org/10.1101/2020.09.28.20203109 (2020).
16. Gao, J. *et al.* STAN: Spatio-temporal attention network for pandemic prediction using real-world evidence. *J. Am. Med. Inform. Assoc.* **28**, 733–743. https://doi.org/10.1093/jamia/ocaa322 (2021).
17. Jin, X., Wang, Y.-X. & Yan, X. Inter-series attention model for COVID-19 forecasting. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 495–503, https://doi.org/10.1137/1.9781611976700.56 (2021).
18. Ray, E. L. *et al.* Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. *medRxiv* https://doi.org/10.1101/2020.08.19.20177493 (2020).

19. Jessi, M. & Luis, F. New York Severely Undercounted Virus Deaths in Nursing Homes, Report Says, Retrieved from https://www.nytimes.com/2021/01/28/nyregion/nursing-home-deaths-cuomo.html (2021).
20. Li, D. *et al.* Identifying US countries with high cumulative COVID-19 burden and their characteristics. *medRxiv* https://doi.org/10.1101/2020.12.02.20234989 (2021).
21. Zhu, S. *et al.* High-resolution Spatio-temporal Model for County-level COVID-19 Activity in the U.S. arXiv:2009.07356 (2020).
22. Chande, A. *et al.* Real-time, interactive website for US-county-level COVID-19 event risk assessment. *Nat. Hum. Behav.* **4**, 1313–1319. https://doi.org/10.1038/s41562-020-01000-9 (2020).
23. Zhang, H., Cissé, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings* (OpenReview.net, 2018).
24. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
25. Harko, T., Lobo, F. S. & Mak, M. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl. Math. Comput.* **236**, 184–194. https://doi.org/10.1016/j.amc.2014.03.030 (2014).
26. Chen, Y.-C., Lu, P.-E., Chang, C.-S. & Liu, T.-H. A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Trans. Netw. Sci. Eng.* **7**, 3279–3294. https://doi.org/10.1109/tnse.2020.3024723 (2020).
27. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
28. Yang, Z. *et al.* Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* **12**, 165 (2020).
29. Xu, C., Yu, Y., Chen, Y. & Lu, Z. Forecast analysis of the epidemics trend of COVID-19 in the USA by a generalized fractional-order SEIR model. *Nonlinear Dyn.* **101**, 1621–1634. https://doi.org/10.1007/s11071-020-05946-3 (2020).
30. Guo, L., Zhao, Y. & Chen, Y. Management strategies and prediction of COVID-19 by a fractional order generalized SEIR model. *medRxiv* https://doi.org/10.1101/2020.06.18.20134916 (2020).
31. Caccavo, D. Chinese and Italian COVID-19 outbreaks can be correctly described by a modified SIRD model. *medRxiv* https://doi.org/10.1101/2020.03.19.20039388 (2020).
32. Baek, J. *et al.* The Limits to Learning a Diffusion Model. arXiv:2006.06373 (2021).
33. Karlen, D. Characterizing the spread of CoViD-19. arXiv:2007.07156 (2020).
34. Sayama, H. *Introduction to the Modeling and Analysis of Complex Systems* (Open SUNY Textbooks, 2015).
35. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734, https://doi.org/10.3115/v1/D14-1179 (Association for Computational Linguistics, 2014).
36. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30 (eds Guyon, I. *et al.*) (Curran Associates, Inc., 2017).
37. Zuo, S., Jiang, H., Li, Z., Zhao, T. & Zha, H. Transformer Hawkes Process. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research* (eds D. III, H. & Singh, A.) 11692–11702 (PMLR, 2020).
38. COVID-19 Modeling. https://bobpagano.com.
39. Microsoft. https://www.microsoft.com/en-us/ai/ai-for-health.
40. Oliver Wyman Pandemic Navigator. https://pandemicnavigator.oliverwyman.com/.
41. CMU Delphi Group. https://delphi.cmu.edu/.
42. Los Alamos National Laboratory. https://covid-19.bsvgateway.org/.
43. University of Massachusetts–Mechanistic Bayesian model. https://github.com/dsheldon/covid.
44. Wang, L. *et al.* Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States. arXiv:2004.14103 (2020).
45. MOBS lab Analysis of the COVID-19 Epidemic. https://www.mobs-lab.org/2019ncov.html.
46. Srivastava, A., Xu, T. & Prasanna, V. K. Fast and Accurate Forecasting of COVID-19 Deaths Using the SIkJα Model arXiv:22007.05180 (2020).
47. Lega, J. Parameter estimation from ICC curves. *J. Biol. Dyn.* **15**, 195–212 (2021).
48. Wu, D. *et al.* DeepGLEAM: a hybrid mechanistic and deep learning model for COVID-19 forecasting. *CoRR* arXiv:2102.06684 (2021).
49. UGA-CEID. https://github.com/CEIDatUGA/COVID-stochastic-fitting.
50. London School of Hygiene and Tropical Medicine. https://www.lshtm.ac.uk/research/centres/centre-mathematical-modelling-infectious-diseases/covid-19.
51. Steve McConnell CovidComplete. https://stevemcconnell.com/covidcomplete/.
52. Zachreson, C. *et al.* Risk mapping for COVID-19 outbreaks in Australia using mobility data. *J. R. Soc. Interface* **18**, 20200657 (2021).
53. James, N. & Menzies, M. Efficiency of communities and financial markets during the 2020 pandemic. arXiv:2104.02318 (2021).
54. Chicchi, L., Giambagli, L., Buffoni, L. & Fanelli, D. Mobility-based prediction of SARS-CoV-2 spreading. arXiv:2102.08253 (2021).
55. Gösgens, M. *et al.* Trade-offs between mobility restrictions and transmission of SARS-CoV-2. *J. R. Soc. Interface* **18**, 20200936 (2021).
56. Carroll, C. *et al.* Time dynamics of COVID-19. *Sci. Rep.* **10**, 21040. https://doi.org/10.1038/s41598-020-77709-4 (2020).
57. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Short Papers)*, 464–468, https://doi.org/10.18653/v1/N18-2074 (Association for Computational Linguistics, 2018).
58. Guo, H., Mao, Y. & Zhang, R. Augmenting Data with Mixup for Sentence Classification: An Empirical Study. *CoRR* arXiv:1905.08941 (2019).
59. Verma, V. *et al.* Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, (eds Chaudhuri, K. & Salakhutdinov, R.) 6438–6447 (PMLR, 2019).
60. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2015).

Author contributions

S.E., S.Y., and T.Z. designed the research; S.E., S.Y., and T.Z. performed the research; S.E. analyzed data; and S.E., S.Y., and T.Z. wrote the paper. S.Y. and T.Z. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-93545-6.

Correspondence and requests for materials should be addressed to S.Y. or T.Z.