Inter-Series Attention Model for COVID-19 Forecasting

Xiaoyong Jin*   Yu-Xiang Wang*   Xifeng Yan*

Abstract

COVID-19 pandemic has an unprecedented impact all over the world since early 2020. During this public health crisis, reliable forecasting of the disease becomes critical for resource allocation and administrative planning. The results from compartmental models such as SIR and SEIR are popularly referred by CDC and news media. With more and more COVID-19 data becoming available, we examine the following question: Can a direct data-driven approach without modeling the disease spreading dynamics outperform the well referred compartmental models and their variants? In this paper, we show the possibility. It is observed that as COVID-19 spreads at different speed and scale in different geographic regions, it is highly likely that similar progression patterns are shared among these regions within different time periods. This intuition lead us to develop a new neural forecasting model, called Attention Crossing Time Series (**ACTS**), that makes forecasts via comparing patterns across time series obtained from multiple regions. The attention mechanism originally developed for natural language processing can be leveraged and generalized to materialize this idea. Among 13 out of 18 testings including forecasting newly confirmed cases, hospitalizations and deaths, **ACTS** outperforms all the leading COVID-19 forecasters highlighted by CDC.

keywords: COVID-19, Time Series Forecasting, Attention, Detrending

1   Introduction

The Coronavirus disease 2019 (COVID-19) has been impacting the human society since early 2020. At the time of this writing, it is an ongoing public health crisis in over 187 countries and territories around the world, with more than 30 million confirmed cases, and a growing death toll exceeding $1,000,000$. During this crisis, reliable forecasting of COVID-19 cases becomes important as it will help (1) healthcare institutes to allocate sufficient supply and resources, (2) policy-makers to consider new and further administrative interventions, (3) general public to be aware of the situation and to follow rules against the epidemic. Therefore, the Center for Disease Control and Prevention (CDC) has been actively collecting and publishing data about confirmed cases, hospitalization and deaths related to COVID-19, and hosting forecasting results in the coming weeks.

The US has been suffering the most severe loss from the pandemic, in which more than $200,000$ lives were lost. To encourage and to bring together efforts of COVID-19 modeling, CDC has launched a forecasting challenge[1]. It calls for models that give predictions of the next 4 weeks on a daily or weekly basis. Besides COVID-19 data, other kinds of data such as demographic data, mobility data and intervention policies are also encouraged to be used in predictions.

Epidemic forecasting is regarded as a challenging task for a long time, for which many methods have been developed. They can be roughly categorized into two classes:

1. **Compartmental models** These models explicitly compartmentalize the population in groups based on their status of infection and recovery, and simulate the transmission process using differential equations. As of today, most of the CDC-featured forecasting methods fall into this category. Examples includes [2, 22, 34] that are built upon classic SIR or SEIR models [11]. Compartmental models describe disease spreading dynamics; however, it is quite hard to determine parameters in these models as they are influenced by many uncontrollable and dynamically changing factors.

2. **Statistical models** This type of methods fits the data to regression models directly, such as [1, 21, 31]. While they are more flexible in processing real data compared to compartmental models, they often assume a simplified model class such as generalized linear models [1], or require sophisticated hand-crafted features from additional, and possibly proprietary, data sources [31].

The forecasting of COVID-19 is even harder as various constantly changing factors, such as virus characteristics, social and cultural distinctions, public attitudes and behaviors, intervention policies and healthcare preparation, influence the contagious rate and death rate significantly. Will there be a better alternative that is solely data-driven without any assumptions about the underlying disease propagation mechanisms? In particular, we experimented a set leading neural forecasters [17, 18, 25], but none of them gave the best result.

---
*University of California, Santa Barbara, CA, USA. Email:{x_jin,xyan,yuxiangw}@cs.ucsb.edu

---
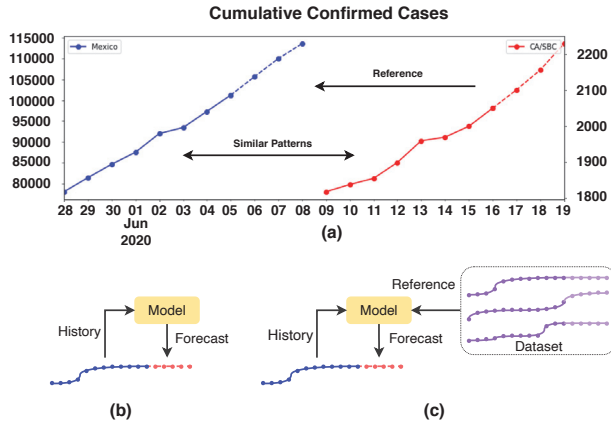[1]https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasting.html

Figure 1: **(a)** A similar growth pattern of confirmed cases in Santa Barbara County, California, in mid June is observed in Mexico in late May and early June. **(b)** Conventional auto-regressive forecasting model. **(c)** The proposed inter-series forecasting model (**ACTS**).

Since the deep models are originally designed for sufficiently long time series with hundreds of points, the scarce historical data in this task might be the reason of their failures. A natural alternative is to exploit other time series in the dataset if they reveals similar dynamics. Fortunately, even if any two regions present different disease curves over long term, it is likely to find short periods in which different regions sharing similar patterns. Figure 1(a) shows surprisingly that the growth pattern of confirmed cases in Santa Barbara county, California, is highly similar to that in Mexico 11 days ago even though at different scales. Moreover, the further growth in Santa Barbara is also close to that within the corresponding time period in Mexico. In light of this key observation, it is intuitively possible to do better forecasting for Santa Barbara by referring to Mexico in this specific time window via proper transformations.

Based on this intuition, we propose to generalize the conventional auto-regressive forecasting to a novel paradigm: besides the local historical data, we also refer to the past reports in all other regions simultaneously in forecasting. Figures 1(b) and (c) illustrate the the fundamental difference between the two paradigms. With time series data of COVID-19 from various locations accumulating over time, we are able to deliver a model outperforming the existing methods by inter-series modeling. Note that unlike other cross-location epidemic forecasters such as [8], only certain time periods rather than the entire time series from other regions will be referred to.

In order to make the proposed paradigm work,

it is critical to find small segments in reference time series that exhibit similarity with target time series. It turns out that the attention mechanism originated in natural language process [29] is a good choice for pattern matching. Moreover, it is found that solely applying attention does not work the best as the embedded small segments do not contain long-term trends that are not directly comparable. We filter out these trends and introduce a normalization step so that the small segments can be matched at a consistent scale. In the end, we put all of these components together and achieve global optimum by joint training. Our new model called **ACTS** (Attention Crossing multiple Time Series), is able to outperform leading forecasters hosted at CDC.

Our main contributions are summarized as follows:

- We develop a new paradigm that leverages inter-series similarity to improve COVID-19 forecasting. Our method makes no assumption about epidemiological dynamics.

- We extend the attention mechanism to capture inter-series similarity in time series data. Trend filtering is also introduced to complement the attention-based framework and it can be trained jointly to maximize the performance.

- In comparison with a wide range of existing forecasters, the outstanding performance of **ACTS** is demonstrated on COVID-19 data.

2 Related Work

There has been a large body of work focusing on epidemic forecasting. To incorporate domain knowledge, mechanistic models [15, 16, 35] has been favored since they often consider various factors such as epidemiological and social properties, and they make forecasts based on simulation. Moreover, geographic information can also be incorporated into the mechanistic models to better illustrate the spreading process of an infectious disease [3, 4]. These models have excellent interpretability but often fail to fit real observed data due to their rigid and over-simplified assumptions without careful calibration.

On the other hand, statistical methods explicitly fit historical data to a statistical model and use it to obtain predictions by extrapolation [5, 6]. For example, [24] relies on kernel density estimation, [19] uses seasonal ARIMA, [33] chooses particle filtering and [36] employs Gaussian process regression. These methods are either too simple or require laborious feature engineering. Hence, various deep learning techniques are also introduced to forecast disease spreading, such as

symbol	interpretation
$\boldsymbol{x}_t^i$	The value at time t in location i.
$\boldsymbol{x}_{s:t}^i$	The time series from s to t in location i
$\boldsymbol{W}$.	Parameter matrices to be learned.
$[\boldsymbol{a}; \boldsymbol{b}]$	The concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$.
$\langle \boldsymbol{a}, \boldsymbol{b} \rangle$	Inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$.
$[\![s,t]\!]$	Consecutive index set $s, s+1, \cdots, t$

Table 1: Used notations

[32, 30, 8, 7, 13, 27, 28, 23, 9]. They use deep neural networks to extract complex temporal patterns from historical data and a selected set of additional features. [8, 9] are conceptually closer to our model, both of which employ attention mechanism to compare encoded temporal patterns across multiple locations. However, they require a fixed graph structure with geographic information and produce a similarity score between locations that is independent of time. Instead, in our model we generate embeddings of dynamical patterns for attention over both spatial and temporal dimensions so that the generated attention map are temporally dynamical and free from any predefined geographic structures.

3 Problem Statement

In COVID-19 forecasting, there are three types of incidences, namely confirmed cases, hospitalizations and deaths, to be predicted. The historical data is reported on a daily basis, and we will predict them for the coming weeks. Table 1 summarizes the notations we use in the following sections. Note that throughout the paper, terms "location" and "region" will be used interchangeably. Problem definition is formulated as follows.

DEFINITION 1. **Incidence Time Series** *We denote by $\boldsymbol{x}_t^i$ the reported value of a certain type of incidence data at date t and location i, for $t = 1, 2, \cdots, T$ and $i = 1, 2, \cdots, N$. Hence, the incidence time series of location i denoted by $\boldsymbol{x}_{1:T}^i$. $\boldsymbol{x}_{s:t}^i$ is called a time segment of $\boldsymbol{x}^i$, where $[\![s,t]\!], 1 \leq s < t \leq T$ is called a window.*

DEFINITION 2. **Target Region** *At the last date T, we predict the future incidences for location $i_0 \in [1, N]$ beyond T. We call i_0 the target region and $\boldsymbol{x}_{1:T}^{i_0}$ the target time series.*

DEFINITION 3. **Reference Regions** *The regions other than the target region i_0 are called reference regions. The reference time series are $\boldsymbol{x}_{1:T}^i$ where $i \neq i_0$. In a generalized definition, reference regions could include the target region.*

DEFINITION 4. **Additional Features** *Besides his-*

torical incidences in each region, other features might be available including demographic information, mobility index, and interventions. For each region i, time-independent features are concatenated into a single vector $\boldsymbol{u}^i$, and time-dependent ones into another time series $\boldsymbol{r}_t^i$.

Problem Statement Given N time series $X_{1:T}^i$ ($i \in [1, N]$) and additional features, we aim to predict future incidences in a target region $i_0 \in [1, N]$ over H consecutive days after T, i.e. $\boldsymbol{x}_{T+1:T+H}^{i_0}$.

4 Methodology

Traditionally, epidemic forecasts are made by analyzing only the growth pattern of incidences. [3, 4, 8, 9] take the incidences from neighboring regions into consideration as diseases spread through social interaction. Rather than explicitly modeling the disease spreading process, we take a bold step to directly compare the incidence curves across regions. Once the similarities between the current incidences in the target region with the past time segments in reference regions are identified, the following incidences in the reference regions can be used to forecast the future incidences in the target region. Hence, the critical challenge in implementing our idea is to (1) define a representation of a time segment; (2) identify similar segments in reference regions through the representations; and (3) aggregate their following incidences for forecasting.

Formally, we introduce an embedding function $\phi(\cdot)$ to encode a time series segment $\boldsymbol{x}_{t-l+1:t}$ into a vector, and then use dot-product of vectors to measure similarity. The following incidences $\boldsymbol{x}_{t+1:t+h}$ is also encoded by another embedding function $\psi(\cdot)$ for further aggregation. However, while there are comparable short-term patterns that can be extracted from time series segments, there are also non-stationary long-term trends that hinder reasonable comparison and aggregation of local patterns within segments.

We resolve the problem in two steps. First, we apply a trainable detrending module to the raw time series to remove long-term trends so that incidences across different regions are more comparable. Second, we take rolling windows from residual time series and transform them into a common feature space using normalized convolution as embedding functions $\phi(\cdot)$ and $\psi(\cdot)$. The embedding of the recent window in the target region is then compared with windows from references to produce weights for combining the following incidences of each reference window. In such pairwise comparisons, differences in both time-dependent and time-independent features are taken into account so that the curves in corresponding windows can be better aligned. The combi-
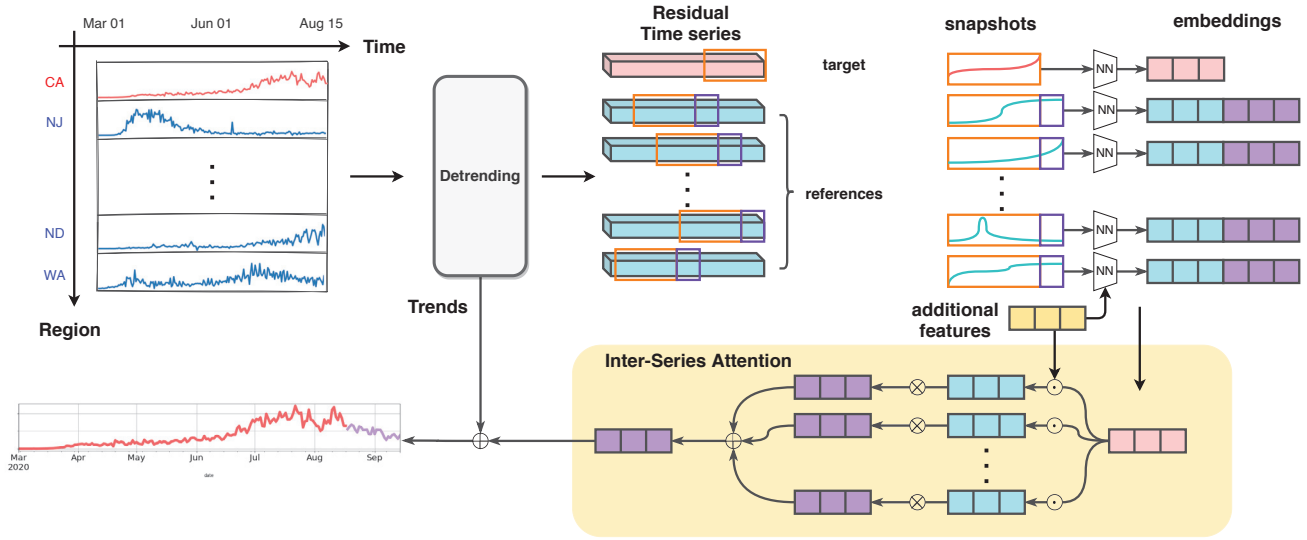
Figure 2: Our proposed Inter-series Attention Network. Best view in color.

nations are then added to the extrapolation of filtered trends to generate the final prediction. We jointly train both modules in an end-to-end manner so that both the long- and short-term patterns can be decoupled in an adaptive way.

Figure 2 gives an overview of the framework. In the following subsections, we introduce each component in details.

4.1   Detrending We adopt a learnable Holt smoothing model ([26]) to remove long-term trends from the raw time series. Specifically, we introduce a set of parameters $\theta_e^i = [a_0^i; b_0^i; \alpha^i; \beta^i]$ per series, where a_0^i is the initial level, b_0^i is the initial trend, α^i is the level smoothing coefficient and β^i is the trend smoothing coefficient. Then Holt's equations ([12]) are launched to iteratively derive levels and piecewise linear slopes in $\boldsymbol{x}_{1:T}^i$,

$$
\begin{aligned}
a_t^i &= \alpha^i \boldsymbol{x}_t^i + (1-\alpha^i)(a_{t-1}^i + b_{t-1}^i), \\
b_t^i &= \beta^i(a_t^i - a_{t-1}^i) + (1-\beta^i)b_{t-1}^i, \\
\hat{\boldsymbol{x}}_t^i &= \boldsymbol{x}_t^i - a_t^i.
\end{aligned}
\tag{1}
$$

After detrending, the residual time series $\hat{\boldsymbol{x}}_{1:T}$ will contain short-term patterns for further processing. Projection from the long-term trend is generated by simple linear extrapolation,

$$
\bar{\boldsymbol{x}}_{t+h}^i = a_t^i + h b_t^i.
\tag{2}
$$

A more sophisticated detrending process might further boost performance; we leave it for future study. The detrending process is applied to all the time series and the residual time series are fed into the following attention module.

4.2   Attention Module As COVID-19 is a new disease, we do not have its historical data in the past seasons. Hence, it is critical to leverage limited data from the same season, but across different regions, i.e. model the correlations between regions that have been undergoing the pandemic. Without detailed information about spatial dynamics such as population movement, we instead employ attention mechanism to measure the relation of one region to other regions by directly comparing the incidence curves after trend filtering. Since there are many stages in a dynamical epidemiological process, it is necessary to learn a representation for each time period in a region for alignment in attention. In light of this idea, we apply a convolution layer to encode the residual time series segment $\hat{\boldsymbol{x}}_{t-l+1:t}$ to a vector, based on which attention scores measuring similarity between regions are computed.

4.2.1   Segment Embedding Even after detrending, the scales of reported numbers in residual time series are still quite different across regions. It is important to normalize residuals before embedding. We empirically find it better to apply min-max normalization to the cumulative sum of incidence time series, which can be regarded as a kind of smoothing. Specifically, for a rolling window of size l representing a period of time, i.e. $\hat{\boldsymbol{x}}_{t-l+1:t}^i, t \in [l, T]$, we compute its cumulative sums and apply the min-max normalization to the monotonically increasing series,

$$
\boldsymbol{c}_j^i = \sum_{k=t-l+1}^{j} \hat{\boldsymbol{x}}_k^i; \qquad \tilde{\boldsymbol{c}}_j^i = \frac{\boldsymbol{c}_j^i - \boldsymbol{c}_{t-l+1}^i}{\boldsymbol{c}_t^i - \boldsymbol{c}_{t-l+1}^i},
\tag{3}
$$

for $j \in [t - l + 1, t]$. As a result, the first and last values of the normalized series will consistently be 0 and 1 respectively.

We then instantiate the function $\phi(\cdot)$ using a convolution layer with d feature maps to the scaled segment and time-dependent features. The kernel size is empirically selected, and when it is smaller than l, average pooling is applied in order to reduce a sequence to a vectorized embedding,

$$(4) \quad \boldsymbol{p}_t^i = \text{AvgPool}\left(\text{Conv}\left([\tilde{\boldsymbol{c}}_{t-l+1:t}^i; \boldsymbol{r}_{t-l+1:t}^i]\right)\right) \in \mathbb{R}^d.$$

These segment embeddings are used to model similarity in different temporal periods across different regions.

Likewise, we employ another convolution-pooling layer as $\psi(\cdot)$ to encode the following incidences over H days after each segment into so-called **development embedding**,

$$(5) \qquad \boldsymbol{g}_t^i = \text{AvgPool}\left(\text{Conv}\left(\tilde{\boldsymbol{c}}_{t+1:t+H}^i\right)\right) \in \mathbb{R}^d.$$

They represent the succeeding development after encoded segments and will be the references for the prediction of the given target region. In fact, we can pair the segments and references by aligning the time indices, i.e. $\{\boldsymbol{p}_t^i, \boldsymbol{g}_t^i\}$ for $t \in [l, T - H]$.

4.2.2 Inter-series Attention Given the embeddings, we use dot-product attention to compare segments and combine the values. Specifically, we linearly map the segment embeddings to query vectors $\boldsymbol{q}_t^i$ and key vectors $\boldsymbol{k}_i^t$, from which the similarity score is computed. The development embeddings are projected to value vectors $\boldsymbol{v}_t^i$. On the other hand, the additional time-independent features $\boldsymbol{u}^i$ are also incorprated into queries and keys.

$$
(6) \qquad
\begin{aligned}
\boldsymbol{q}_t^i &= \boldsymbol{W}_Q \boldsymbol{p}_t^i + \boldsymbol{W}_{u,q} \boldsymbol{u}^i; \\
\boldsymbol{k}_t^i &= \boldsymbol{W}_K \boldsymbol{p}_t^i + \boldsymbol{W}_{u,k} \boldsymbol{u}^i; \\
\boldsymbol{v}_t^i &= \boldsymbol{W}_V \boldsymbol{g}_t^i;
\end{aligned}
$$

For a target region i_0, we take $\boldsymbol{q}_T^{i_0}$ for the last segment and compute its similarity with all the keys from other time segments across all the regions, which is then used to obtain a weighted sum of values.

$$(7) \qquad \hat{\boldsymbol{v}}_T^{i_0} = \sum_{i,t \in \Omega} \frac{\exp\left(\langle \boldsymbol{q}_T^{i_0}, \boldsymbol{k}_t^i \rangle\right)}{\sum_{i,t \in \Omega} \exp\left(\langle \boldsymbol{q}_T^{i_0}, \boldsymbol{k}_t^i \rangle\right)} \boldsymbol{v}_t^i,$$

where $\Omega = [1, N] \times [l, T - H]$. In this way, the past observations in both the target region and reference regions are fully utilized. The weighted combination of values $\hat{\boldsymbol{v}}_T^{i_0}$ is then linearly projected to an estimate of

$\tilde{\boldsymbol{c}}_{T+1:T+H}^{i_0}$. We apply the inverse transformation of (3) to get an estimate of $\hat{\boldsymbol{x}}_{T+1:T+H}^{i_0}$, denoted by $\hat{\boldsymbol{y}}_{T+1:H}^{i_0}$.

In the end, the estimate from attention module is added to the extrapolations in the detrending module to produce the final forecast $\boldsymbol{y}_{T+1:T+H}^{i_0}$, where

$$\boldsymbol{y}_t^{i_0} = \bar{\boldsymbol{x}}_t^{i_0} + \hat{\boldsymbol{y}}_t^{i_0}, \qquad t \in [T+1, T+H].$$

4.3 Joint Training The model can be trained by minimizing the joint loss with respect to the parameters in all the modules. The joint loss is an aggregation of prediction error $E(\cdot, \cdot)$ computed in two steps. First, for a single target region, we compare our forecasts and ground truths for different T, i.e. lengths of history. Second, we take the aggregated loss in the first step for every region. Formally, the joint loss is defined as

$$(8) \qquad \mathcal{L} = \sum_{i=1}^{N} \sum_{T=l}^{L-H} E(\boldsymbol{y}_{T+1:T+H}^i, \boldsymbol{x}_{T+1:T+H}^i)$$

where L is the total number of available historical reports, and l is the minimum required history length. In our experiments, we choose Mean Absolute Error (MAE) to be the error metric $E(\cdot, \cdot)$, i.e.

$$E(\boldsymbol{y}_{T+1:T+H}^i, \boldsymbol{x}_{T+1:T+H}^i) = \frac{1}{H} \sum_{t=T+1}^{T+H} |\boldsymbol{y}_t^i - \boldsymbol{x}_t^i|$$

5 Experiments

In this section, we demonstrate the effectiveness of the proposed model on real COVID-19 datasets. We intend to answer the following questions:

- Can **ACTS** outperform the popular COVID-19 forecasters referred at CDC and other state-of-the-art deep learning models?

- How much does each component of **ACTS** contribute to the model performance?

- What kind of similarity can inter-series attention capture?

5.1 Experimental Settings
Dataset The COVID-19 incidence data is publicly available at JHU-CSSE[2] and COVID tracking project[3]. Additional features are also publicly available [4] [5] [6].

---

[2]github.com/CSSEGISandData/COVID-19
[3]covidtracking.com/
[4]github.com/descarteslabs/DL-COVID-19
[5]github.com/djsutherland/pummeler
[6]data.world/liz-friedman/hospital-capacity-data-from-hghi

The features we used include total population, population density, ratios of age/gender/race, available hospital beds, and traffic mobility, which are proven to bring marginal accuracy gain in the hospitalization forecasting task in our experiments. The dataset covers the reports up to September 27, 2020 from 50 states and DC in the US.

Evaluation Protocol As required by CDC, we predict the incidence data over the next 4 weeks at a given date and compare the forecasts with the reported ground truths. Suppose we are predicting the new confirmed cases in the state of California starting from 08/16. As context, we are provided a daily time series consisting of incidences in all the states till 08/15. There are three forecasting tasks: daily forecasts for new hospitalizations, weekly forecasts for new confirmed cases and deaths.

The forecasting performance is evaluated in terms of Weighted Absolute Percentage Error (WAPE), defined by the ratio of Mean Absolute Error (MAE) and mean value of ground truths and frequently used in research [17, 18]. At each prediction date, we keep the data in the last 7 days for validation, and the remaining historical data for training. We use the validation data to tune the hyperparameters and to avoid overfitting by early stopping. Other implementation details can be found in Appendix too.

Baselines We compare the performance of the epidemic models featured at CDC, including

- **YYG** [10]: An SEIR model with learnable parameters that attracts a lot of attention from media;

- **CU** [22]: A metapopulation SEIR model developed by researchers in Columbia University;

- **UCLA** [37]: An SuEIR model using machine learning developed by Statistical Machine Learning Lab at UCLA;

- **ERDC**[7]: An SEIR model that considers unreported infections and isolated population developed by US Army Engineer Research and Development Center;

- **LANL** [14]: A statistical dynamical growth model accounting for population susceptibility developed by Los Alamos National Laboratory;

- **CovidSim**[8]: Machine learning model based on generalized random forests.

The first four are compartmental models and the last two rely on statistical modelling. Other than these conventional models, we also evaluate three deep learning models for time series forecasting,

- **DeepCOVID** [25] An operational deep learning framework designed for real-time COVID-19 forecasting developed by Georgia Tech;

- **ConvTrans** [17] A self-attention based Transformer model that also employs convolutions for pattern representations;

- **TFT** [18] A self-attention based deep learning model with feature selection.

We implement the ConvTrans and TFT and tune the hyperparameters using the validation data. All of our implementations run on a server with an Intel i7-6700K CPU and a single GTX 1080Ti GPU. For other baselines, since their implementations are not open-sourced, we take their forecasts submitted to the challenge hosted by CDC [9].

5.2 Performance Comparison Table 2 shows the forecasting performance on 6 different dates. Three types of incidence data, namely confirmed cases (C), hospitalizations (H) and deaths (D) are separately predicted. We have three key observations: (1) In 13 out of 18 cases, **ACTS** outperforms other algorithms by a considerable margin. On average, it improves 9%, 5% and 4% over the best of these algorithms for C, H and D, respectively. (2) **ACTS** is more favorable on recent days when there are more abundant data available, showing that data-driven methods benefit from more data. (3) The two deep learning approaches ConvTrans and TFT do not exhibit strong performance. The main difference between ours and these approaches is the employment of attention across multiple time series, which dramatically boosts the performance. Note that our model can be trained in less than 5 minutes and inference takes only seconds.

5.3 Ablation Study For deeper understanding of our model, we disable each component of **ACTS** to examine its contribution:

- **ACTS-d** We remove the detrending module and obtain an attention-only forecaster;

- **ACTS-n** We remove the normalization in segment embedding;

---

[7]https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/USACE-ERDC_SEIR/metadata-USACE-ERDC_SEIR.txt

[8]https://www.covid19sim.org/documents/outbreak-methods

[9]https://github.com/reichlab/covid19-forecast-hub

		Method									
		YYG	CU	UCLA	ERDC	LANL	Covid Sim	Deep COVID	Conv Trans	TFT	ACTS
06/21	C	-	-	-	-	0.51	-	-	1.09	0.51	**0.39±0.01**
	H	-	1.91	-	-	1.08	0.95	**0.63**	1.22	0.80	0.80±0.02
	D	0.52	1.48	0.56	-	0.58	1.46	0.66	1.09	0.67	**0.45±0.01**
07/05	C	-	-	-	-	0.37	-	-	0.37	0.39	**0.33±0.01**
	H	-	0.98	1.23	0.66	0.95	-	0.65	1.08	0.84	**0.61±0.04**
	D	0.45	0.65	0.53	**0.38**	0.52	-	0.85	0.60	0.51	0.60±0.01
07/19	C	-	-	-	-	**0.27**	-	-	0.50	0.44	0.31±0.01
	H	-	0.67	1.24	0.77	0.78	1.71	0.70	0.99	0.66	**0.60±0.03**
	D	0.30	0.43	0.39	1.10	0.48	0.33	0.4506	0.54	0.67	**0.28±0.01**
08/02	C	-	-	-	-	0.30	-	-	0.24	0.24	**0.16±0.04**
	H	-	0.67	0.95	0.71	0.68	1.66	0.79	0.93	0.92	**0.66±0.09**
	D	0.24	0.37	0.27	0.57	0.44	0.26	0.29	0.45	0.38	**0.21±0.01**
08/16	C	-	0.67	0.35	0.28	0.29	0.23	-	0.33	0.55	**0.20±0.03**
	H	-	0.64	0.99	0.60	0.65	1.38	0.98	0.96	0.92	**0.57±0.02**
	D	**0.19**	0.42	0.25	0.53	0.34	0.27	0.28	0.44	0.31	0.23±0.01
08/30	C	-	0.43	0.31	0.34	0.33	0.23	-	0.36	0.29	**0.23±0.03**
	H	-	0.66	0.91	0.68	0.69	1.31	0.83	0.93	0.82	**0.58±0.03**
	D	**0.20**	0.41	0.23	0.56	0.34	0.25	0.36	0.42	0.40	0.25±0.02

Table 2: Forecasting performances across different time periods for different types of incidence data in terms of WAPE. A smaller value indicates better performance. We also include the variance of our model's performance by running 5 times with different random initalizations. "-" means the forecasting results of the corresponding baseline are not available.
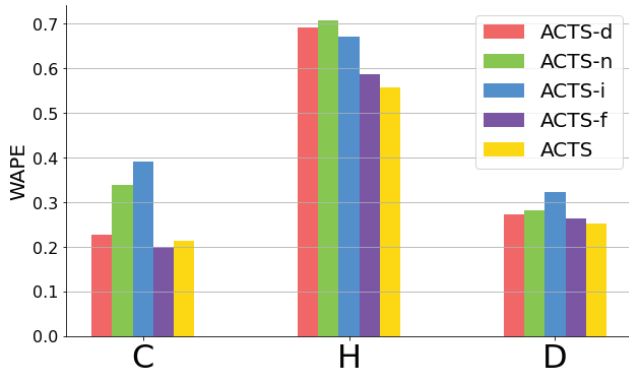

Figure 3: Empirical effects of each component of **ACTS** on forecasting error.

- **ACTS-i** We restrict the attention to the target time series only. The model degenerates to an auto-regressive model similar to ConvTrans and TFT;

- **ACTS-f** We remove the additional features in the model and only rely on incidence data.

The hyperparameters of all variants are kept the same. We compare their performance against **ACTS** using training data up to August 30, 2020. Figure 3 depicts the results, based on which we have the following observations:

- Overall every component of **ACTS** has positive effects on forecasting accuracy, except that the introduction of additional features has mixed effect. We suspect that either better modelling could help or their effect has been absorbed by the incidence time series;

- Among all the components, inter-series attention has the most significant impact on the performance, which proves that our design of attention crossing multiple time series is valid. It can capture cross-region similarity in COVID-19 forecasting;

- The detrending module makes some contribution. We believe it has the potential for further improvement, e.g. employing advanced trend filtering or even epidemic models.
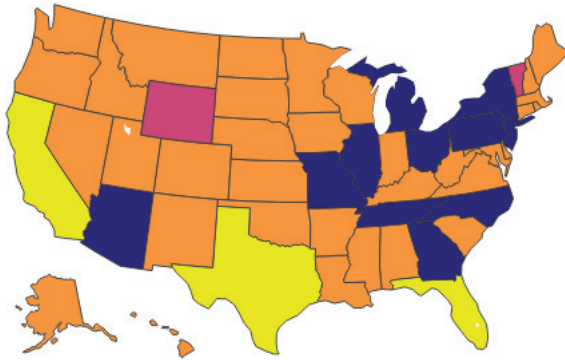
Figure 4: Groups of the US states learned by inter-series attention on death tolls by August 30, 2020.

5.4 Cross-region Similarity A key feature of **ACTS** is that it can capture similarity between regions via attention from data. According to (7), the reference set Ω is common for any target regions i_0, and the learned attention distribution is determined by $\boldsymbol{q}_T^{i_0}$. Hence, we directly take those d-dim queries for every region and apply K-means clustering to group them. In this experiment, we use the death forecasting model as an example, where T is August 30, 2020, and $K = 4$ is selected based on the Elbow method [20].

A colored map is shown in Figure 4 based on obtained clusters. We can see that California, Texas and Florida, the three states recently hit most seriously are grouped together. Furthermore, states like Arizona, Illinois, North Carolina and Georgia are recognized since they also suffer severe crisis. Interestingly, the states of Wyoming and Vermont are distinguished by our model, in which few deaths are observed for a long period. Overall, our method is able to identify similarities between regions to a certain degree.

6 Conclusion

In this paper, we present **ACTS** for COVID-19 forecasting, a purely data-driven framework for an urgent forecasting problem concerning the entire world. It extends the popular deep learning technique, namely attention mechanism, to learning inter-series similarity for time series forecasting. Above that, we also introduce a detrending component to model long-term trends that are difficult for attention model to capture. Both modules are learned jointly based solely on COVID-19 incidence data and a handful of simple features. Without any domain knowledge, our model can empirically outperform many strong forecasters that are featured by CDC. On

the other hand, we find great potential for improvement for trend filtering and incorporating additional features, which is left to future work.

References

[1] ALTIERI, N., BARTER, R. L., DUNCAN, J., DWIVEDI, R., KUMBIER, K., LI, X., NETZORG, R., PARK, B., SINGH, C., TAN, Y. S., ET AL. Curating a covid-19 data repository and forecasting county-level death counts in the united states. *arXiv preprint arXiv:2005.07882* (2020).

[2] ARIK, S. O., LI, C.-L., YOON, J., SINHA, R., EPSHTEYN, A., LE, L. T., MENON, V., SINGH, S., ZHANG, L., YODER, N., ET AL. Interpretable sequence learning for covid-19 forecasting. *arXiv preprint arXiv:2008.00646* (2020).

[3] BALCAN, D., COLIZZA, V., GONÇALVES, B., HU, H., RAMASCO, J. J., AND VESPIGNANI, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21484–21489.

[4] BALCAN, D., GONÇALVES, B., HU, H., RAMASCO, J. J., COLIZZA, V., AND VESPIGNANI, A. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of computational science 1*, 3 (2010), 132–145.

[5] BROOKS, L. C., FARROW, D. C., HYUN, S., TIBSHIRANI, R. J., AND ROSENFELD, R. Flexible modeling of epidemics with an empirical bayes framework. *PLoS Comput Biol 11*, 8 (2015), e1004382.

[6] CHAKRABORTY, P., KHADIVI, P., LEWIS, B., MAHENDIRAN, A., CHEN, J., BUTLER, P., NSOESIE, E. O., MEKARU, S. R., BROWNSTEIN, J. S., MARATHE, M. V., ET AL. Forecasting a moving target: Ensemble models for ili case count predictions. In *Proceedings of the 2014 SIAM international conference on data mining* (2014), SIAM, pp. 262–270.

[7] CHIMMULA, V. K. R., AND ZHANG, L. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals* (2020), 109864.

[8] DENG, S., WANG, S., RANGWALA, H., WANG, L., AND NING, Y. Graph message passing with cross-location attentions for long-term ili prediction. *arXiv preprint arXiv:1912.10202* (2019).

[9] GAO, J., SHARMA, R., QIAN, C., GLASS, L. M., SPAEDER, J., ROMBERG, J., SUN, J., AND XIAO, C. Stan: Spatio-temporal attention network for pandemic prediction using real world evidence. *arXiv preprint arXiv:2008.04215* (2020).

[10] GU, Y. Covid-19 projections using machine learning. https://covid19-projections.com. Accessed: 2020-10-05.

[11] HARKO, T., LOBO, F. S., AND MAK, M. Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal

death and birth rates. *Applied Mathematics and Computation 236* (2014), 184–194.

[12] Holt, C. C. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting 20*, 1 (2004), 5–10.

[13] Huang, C.-J., Shen, Y., Kuo, P.-H., and Chen, Y.-H. Novel spatiotemporal feature extraction parallel deep neural network for forecasting confirmed cases of coronavirus disease 2019. *medRxiv* (2020).

[14] LANL COVID-19. COVID-19 confirmed and forecasted case data. https://covid-19.bsvgateway.org/, 2020. [Online; accessed 29-May-2020].

[15] Lessler, J., Azman, A. S., Grabowski, M. K., Salje, H., and Rodriguez-Barraquer, I. Trends in the mechanistic and dynamic modeling of infectious diseases. *Current Epidemiology Reports 3*, 3 (2016), 212–222.

[16] Lessler, J., and Cummings, D. A. Mechanistic models of infectious disease and their impact on public health. *American journal of epidemiology 183*, 5 (2016), 415–422.

[17] Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems* (2019), pp. 5243–5253.

[18] Lim, B., Arik, S. O., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv preprint arXiv:1912.09363* (2019).

[19] Martinez, E. Z., Silva, E. A. S. d., and Fabbro, A. L. D. A sarima forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil. *Revista da Sociedade Brasileira de Medicina Tropical 44*, 4 (2011), 436–440.

[20] Marutho, D., Handaka, S. H., Wijaya, E., et al. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication* (2018), IEEE, pp. 533–538.

[21] Murray, C., et al. Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries.

[22] Pei, S., and Shaman, J. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv* (2020).

[23] Ramchandani, A., Fan, C., and Mostafavi, A. Deepcovidnet: An interpretable deep learning model for predictive surveillance of covid-19 using heterogeneous features and their interactions. *IEEE Access* (2020).

[24] Ray, E. L., Sakrejda, K., Lauer, S. A., Johansson, M. A., and Reich, N. G. Infectious disease prediction with kernel conditional density estimation. *Statistics in medicine 36*, 30 (2017), 4908–4929.

[25] Rodriguez, A., Tabassum, A., Cui, J., Xie, J., Ho, J., Agarwal, P., Adhikari, B., and Prakash, B. A. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. *medRxiv* (2020).

[26] Smyl, S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting 36*, 1 (2020), 75–85.

[27] Tian, T., Jiang, Y., Zhang, Y., Li, Z., Wang, X., and Zhang, H. Covid-net: A deep learning based and interpretable predication model for the county-wise trajectories of covid-19 in the united states. *medRxiv* (2020).

[28] Tian, Y., Luthra, I., and Zhang, X. Forecasting covid-19 cases using machine learning models. *medRxiv* (2020).

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[30] Wang, L., Chen, J., and Marathe, M. Defsi: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 9607–9612.

[31] Woody, S., Tec, M. G., Dahan, M., Gaither, K., Lachmann, M., Fox, S., Meyers, L. A., and Scott, J. G. Projections for first-wave COVID-19 deaths across the us using social-distancing measures derived from mobile phones. *medRxiv* (2020).

[32] Wu, Y., Yang, Y., Nishiura, H., and Saitoh, M. Deep learning for epidemiological predictions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018), pp. 1085–1088.

[33] Yang, W., Karspeck, A., and Shaman, J. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol 10*, 4 (2014), e1003583.

[34] Yang, W., Shaff, J., and Shaman, J. Covid-19 transmission dynamics and effectiveness of public health interventions in new york city during the 2020 spring pandemic wave. *medRxiv* (2020).

[35] Zhang, Q., Perra, N., Perrotta, D., Tizzoni, M., Paolotti, D., and Vespignani, A. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th international conference on world wide web* (2017), pp. 311–319.

[36] Zimmer, C., and Yaesoubi, R. Influenza forecasting framework based on gaussian processes. 1–10.

[37] Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W., and Gu, Q. Epidemic model guided machine learning for covid-19 forecasts in the united states. *medRxiv* (2020).

A Implementation Details

We implement our model and its variants using Py-Torch. The hyperparameters we used in all of our experiments are listed in the following table.

Hyperparam	values
hidden size d	[16, 32]
segment length l	[7, 14]
horizon H	7
learning rate	[0.001, 0.005, 0.01]
# training iterations	[600, 1200, 1800]

Table 3: Hyperparameters

Exact values are selected by validation loss. We train all of our implementations using an Intel i7-6700K CPU and a single NVIDIA GTX 1080 Ti GPU (CUDA 10.2) hosted by Ubuntu 16.04. Each training iteration takes approximately $\frac{1}{6}$ second.

Since for each task we need to forecast 4 weeks, we separately predict each week using the same attention module to avoid long-term errors. To predict the k-th week, we replace (5) by

$$\boldsymbol{g}_t^i = \mathrm{AvgPool}\left(\mathrm{Conv}\left(\tilde{\boldsymbol{c}}_{t+(k-1)H+1:t+kH}^i\right)\right),$$

i.e. take the development $(k-1)H$ days after the corresponding segment. For case and death forecasting where forecasts are aggregated by weeks, we directly aggregate $\boldsymbol{g}_t^i$ within a week before applying the final transformation.

In generating final prediction, we avoid negative values by clipping partial predictions from both detrending module and attention module.

Our implementation is open-sourced at `https://github.com/Gandor26/covid-open`.

B Example Forecasts

We here show example forecasts of hospitalizations and deaths in three representative states, i.e. Florida, Maryland and Virginia, to give a qualitative demonstration. Our forecasts are accompanied by two best baselines in either task.

We can see that in most cases **ACTS** fits the ground truths better than baselines. An exception is the hospitalization forecast for Maryland, in which **ACTS** has systematical underestimation. This is because the downward trend captured by the detrending module significantly drags the final prediction down. It indicates that more advanced trend filtering method can further improve the performance of our model.
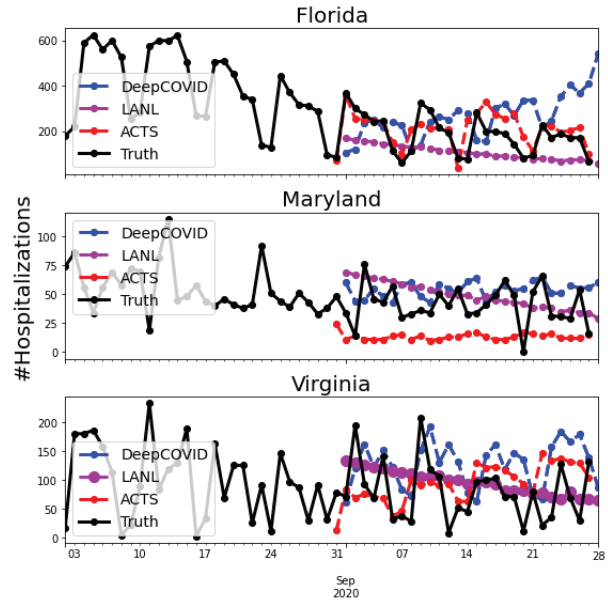

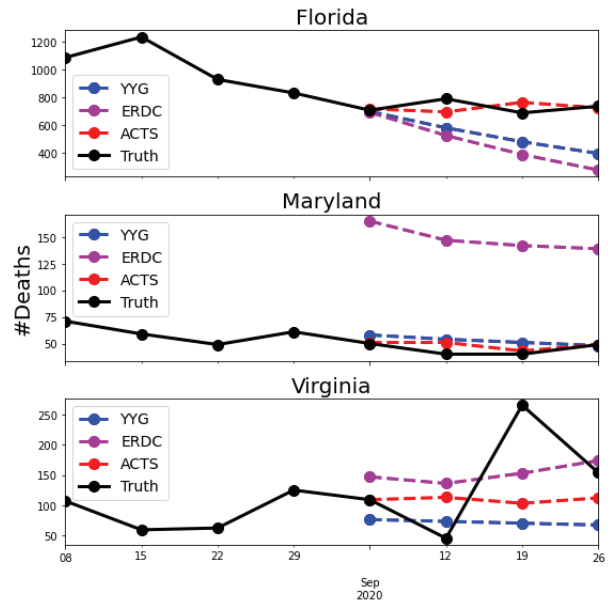Figure 5: Daily hospitalization forecasts on August 30, 2020

Figure 6: Weekly death forecasts on August 30, 2020