Spatio-temporal convolutional residual network for regional commercial vitality prediction

Dongjin Yu1 · Xinfeng Wang1 · Ping Liang1 · Xiaoxiao Sun1

Received: 8 February 2021 / Revised: 28 January 2022 / Accepted: 9 March 2022 / Published online: 29 March 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The vitality of commercial entities reflects the business condition of their surrounding area, the prediction of which helps identify the trend of regional development and make investment decisions. The indicators of business conditions, like revenues and profits, can be employed to make a prediction beyond any doubt. Unfortunately, such figures constitute business secrets and are usually publicly unavailable. Thanks to the rapid growing of location based social networks such as Yelp and Foursquare, massive amount of online data has become available for predicting the vitality of commercial entities. In this paper, a Spatio-Temporal Convolutional Residual Neural Network (STCRNN) is proposed for regional commercial vitality prediction, based on public online data, such as reviews and check-ins from mobile apps. Firstly, a commercial vitality map is built to indicate the popularity of business entities. Afterwards, a local convolutional neural network is employed to capture the spatial relationship of surrounding commercial districts on the vitality map. Then, a 3-dimension convolution is applied to deal with both recent and periodic variations, i.e., the sequential and seasonal changes of commercial vitality. Finally, long short-term memory is introduced to synthesize these two variations. In particular, a residual network is used to eliminate gradient vanishing and exploding, caused by the increase of depth of neural networks. Experiments on public Yelp datasets from 2013 to 2018 demonstrate that STCRNN outperforms the current methods in terms of mean square error.

Keywords Commercial vitality prediction · Commercial district · Spatio-temporal analysis · Convolutional neural network · Residual network · Visual data analysis · Yelp

This work was partially supported by National Natural Science Foundation of China (No.61702144), Industrial Internet Innovation and Development Project of Ministry of Industry and Information Technology (No. TC200802G, No. TC2008033), Key Research and Development Program of Zhejiang Province (No. 2020C01165), Natural Science Foundation of Zhejiang Province (No.LQ20F020017).

Dongjin Yu
yudj@hdu.edu.cn

Extended author information available on the last page of the article.
1 Introduction

The commercial district is a particular area of a town or country that includes clustered business entities like shops, theaters and restaurants, etc. The vitality of these entities in turn reflects the commercial condition or economic strength of their surrounding district. To discover the trend of regional commercial vitality plays a very important role in many potential applications. For example, if a commercial district is believed to keep flourishing or booming, it is worth investing in. Otherwise, if believed to fall in a recession in future, the surrounding house price will become declining. Another example of potential application comes from urban planning. Urban planners decide how facilities should be equipped if a business circle keeps growing. If there will be a large number of guests, more convenience stores need to open. Under all such circumstances, the decision makers are required to have a clear insight of how the commercial district evolves in a particular region. However, to accurately predict the trend of the regional commercial vitality is not a straightforward task. One of the difficulties comes from the lack of data for prediction. It is generally known that the business performance, such as the revenue and profit of individual entities, is commercial secret and not always publicly available.

As a matter of fact, the rise of social media and other digital technologies offers new opportunities to study the perception of urban environments [13]. Meanwhile, during the past few years, the concepts of big data and smart city have become increasing popular, which introduced a new means of predicting commercial vitality, based on the huge quantities of data, collected from different aspects of cities. On the other hand, with the growth of Yelp and other online social apps, a large number of reviews and check-in records of various commercial entities, provided by visitors and consumers, have become available, making it possible to judge the commercial vitality of each area. In other words, the greater the number of reviews and check-in records, the more active the corresponding area can be.

On the other hand, from a technical perspective, the time series prediction methods, such as Auto-Regressive Integrated Moving Average (ARIMA), have been widely applied in the commercial prediction field [35], which, however, usually fail to capture the complex, non-linear, spatio-temporal variations of commercial vitality as a geographic phenomenon. Besides, certain other existing works ignore the periodic time relationships and include irrelevant spatial relationships among remote entities [11]. Thus, they cannot model the complex, non-linear, spatio-temporal relationship of commercial evolvement, resulting in unsatisfactory prediction performance.

To address the aforementioned problems, this paper proposes a novel model, called Spatio-Temporal Convolutional Residual Neural Network (STCRNN), to predict the future commercial vitality, based on public online data from mobile apps, such as reviews and check-in records. Generally speaking, more reviews and check-ins indicate more vitality in the corresponding district. In other words, the number of reviews and check-ins can reflect the commercial vitality. As an example, the commercial vitality of a city, for instance Las Vegas, can be characterized by online reviews and check-ins in relation to the business entities located in the city, from social apps such as Yelp. In this way, a commercial vitality map of Las Vegas can be built, with the vitality on each grid being represented by a color value. On the map, the increased level of vitality in the district around Bellagio Fountains is probably indicated by a bright color, whereas certain remote areas with low levels of vitality are represented by a darker color. The task is to investigate the distribution of regional commercial vitality and its variation tendency, by predicting how the commercial vitality map will change in the future.
The presented model is motivated by the outstanding performance of deep learning techniques on handling non-linear relations, more specifically in terms of predicting spatio-temporal scenarios such as air pollution [12] and taxi demand scheduling [32]. According to the first law of geography “Near things are more related than distant things” [26], therefore, the neighboring commercial entities in the same district are more likely to affect the commercial vitality of one other. In order to capture these spatial relationships, the local Convolutional Neural Networks (CNN) [14] is applied to commercial vitality prediction, which ensures that close spatial relationships are identified as most significant. Further, through the analysis of the Yelp datasets, the reviews and check-in records demonstrate the periodic changes in certain commercial entities. For example, the number of reviews and check-ins of ice cream stores increases significantly in summer rather than in winter; more customers visit shopping malls during holidays than on workdays. In order to reflect this reality, a 3D convolution [27] is adopted to extract the recent variations and periodic rules, and a Long Short-Term Memory (LSTM) model [4] is employed to synthesize them as the temporal characteristics for commercial vitality prediction. Experiments on public Yelp datasets demonstrate that STCRNN outperforms the current methods in terms of Mean Square Error (MSE).

Some existing methods utilize the clustering techniques and visualization tools to empirically explore the commercial activeness [31], or primary neural networks, i.e., fully connected-based neural network [28], to predict the commercial vitality. Few of them can sufficiently model vitality variations with time by exploiting the multiple contexts, i.e., spatial, temporal and periodic contexts from check-in dataset. To make matters worse, many of them highly depend on personal experiences, thus leading to huge analysis cost. By contrast, STCRNN addresses the above issues by learning the spatial-temporal and periodic representations to enhance the prediction performance. In summary, the main contributions of this paper are summarized as follows:

– It is the first comprehensive approach to apply the deep learning technology to predict the vitality of commercial districts based on publicly available online reviews and the check-in records of commercial entities from mobile social apps.
– A spatial dimension is designed to employ a local CNN to capture the spatial relationships, particularly, of neighboring commercial districts, but remove the irrelevant influence from distant commercial districts.
– A temporal dimension is introduced to apply 3D convolutions to extract the periodic and recent variations of commercial vitality, and LSTM is employed to further quantify their contributions of temporal evolution.

The remainder of the paper is structured as follows. After discussing related work in Section 2, the preliminaries and the problem definition are introduced in Section 3. Section 4 presents the prediction model in detail. The experimental results and their discussion are provided in Sections 5 and 6 respectively. Finally, Section 7 concludes the paper and outlines the future work.

2 Related works

Space and time are two fundamental dimensions related to all geographic research. For a long time, spatio-temporal analysis and modeling of geographic parameters have been the
main focus of Geographic Information Science (GIScience) for applications such as predicting urban growth [5], coastal sea variation [6], land use change [29], and urban water quality [18]. Commercial vitality, which changes with the aggregation and evolution of commercial districts, is also a geographic issue affected by complex space and time factors. Traditional studies on this issue are generally carried out by investigating local conditions, which is unsustainable and relies heavily on field surveys [19]. Recent studies attempt to explore the utilities of big data, such as social and review data generated from mobile apps (e.g., Twitter, Yelp, etc.), in commercial vitality prediction. For example, Yang et al. [31] apply a clustering algorithm to aggregate commercial districts based on multiple online data and employ a linear model to predict commercial vitality. The linear-based method, however, does not have the sufficient capability of exploiting the complex spatial-temporal data. Xu et al. [30] present an analytical framework to unravel the landscape and pulses of cycling activities from a dockless bike-sharing system by using a four-month Global Positioning System (GPS) dataset collected from a major bike-sharing operator in Singapore. Yuan et al. [34] propose an approach to discovering city regions with different functions using human mobility and Point-of-Interests (POIs). Both works directly utilize the shared-bike distribution only to discover the activities ranges without any predicting schemes. Wang et al. [28] predict business failure with mobile, location-based check-ins, but ignore the multiple complex contexts. In summary, all the above mentioned studies essentially do not recognize the variations of commercial vitality as a non-linear spatio-temporal issue, which leads to relatively low accuracy in terms of the prediction results. It is noteworthy that it could be a possible way to handle the huge amount of social data by leveraging the cloud computing. Under such circumstance, to ensure data privacy must be given high priority [20].

Deep learning, vastly applied in the field of image and video processing, has been identified as being able to handle complex, non-linear relationships in spatio-temporal prediction. For example, He et al. [10] put forward a multi-view ensemble neural network to predict commercial hotness. In certain sub-neural networks, they introduce CNN. Zhang et al. [36] present a deep learning model with 2D CNN to predict urban congestion. More recently, Guo et al. [9] employ the deep spatio-temporal 3D CNN for traffic forecasting. In these studies, however, the entire research area is fed into the CNN as one image, which fails to capture the local relationships among the surrounding areas, and falsely includes the irrelevant relationships of remote entities. In addition, Ji [14] and Tran [27] et al. demonstrate that 3D convolution can perceive not only spatial but also temporal features compared with 2D convolution in the field of video analysis. Another technology worthy of note is LSTM, which has recently been successfully applied to solving spatio-temporal issues due to its outstanding ability in capturing temporal relationships. For example, Chen et al. [4] apply LSTM to forecast urban housing price and Kong et al. [15] utilize LSTM to forecast urban power load. Since LSTM cannot reflect the spatial relationships, researchers think that combining CNN and LSTM may capture both spatial and temporal characteristics. For example, Huang et al. [12] apply the CNN-LSTM model to predict air particulate matter (PM2.5). Li et al. [17] combine CNN and LSTM to predict the travel distance and Origin-Destination distribution of shared bicycles under different conditions of time and space. However, due to the complex model architecture, the depths of the deep learning network increase sharply, which leads to gradient vanishing and exploding and finally reduces the effectiveness of capturing spatio-temporal relationships [11]. On the other hand, LSTM alone cannot capture both the periodic temporal relationships (e.g., seasonal changes or holiday effects) in spatio-temporal modeling, which is of vital importance in long-term prediction [36].
order to settle spatio-temporal sequence forecasting problems, Shi et al. [23] propose Convolutional LSTM networks (ConvLSTMs). However, due to their complex architectures, training becomes more difficult when the networks’ depths increase, which, in turn, limits their capabilities of capturing wide-ranged, spatio-temporal correlation. Besides, as the authors’ previous work, in [33] a very rough idea of employing deep learning model for commercial activeness prediction is demonstrated.

Table 1 summarizes the major findings from the existing schemes, and their advantage and disadvantage.

In conclusion, the prediction of regional commercial vitality is a complex and non-linear geographic issue, containing both spatial and temporal variations. However, to date, these two characteristics have not been adequately extracted at the same time. Furthermore, the temporal features of regional commercial vitality should be considered from both periodic and recent dimensions. Unfortunately, existing methods, even those based on deep learning, however, cannot capture these two dimensions simultaneously and effectively. Therefore, this paper is dedicated to proposing a deep learning model based on local CNN, 3D convolution, LSTM and residual networks for regional commercial vitality prediction.

3 Preliminaries

This section presents the definitions and the background knowledge as the preliminaries in order to understand the proposed model.

3.1 Definitions

**Definition 1** (Commercial entity) A business entity τ, such as a restaurant, shop and theater is presented as (τ.id, τ.la, τ.lo), in which τ.id represents the id of each business entity and τ.la and τ.lo represent the latitude and longitude of each business entity, respectively.

**Definition 2** (Commercial district and commercial grid) The aggregation of commercial entities locates in a neighboring space forms a commercial district and is indicated by \( D = \{\tau_0, \tau_2, \ldots, \tau_n\} \). A commercial district can also be divided into a series of commercial grids, indicated by \( g_{i,j} \) or \((i, j)\). One commercial grid may have the diffusion and radiation effect on its surrounding grids.

**Definition 3** (Commercial vitality) The number of reviews and check-ins related to one commercial entity \( \tau \), within a certain time slot \( t \), is used to represent its commercial vitality, denoted by (1).

\[
y^\tau_t = \sum_{i \in \tau, t} |\text{Review}_i| + \sum_{i \in \tau, t} |\text{Checkin}_i| \quad (1)
\]

Similarly, the commercial vitality of one commercial grid \( g \), within a certain time slot \( t \), is the sum of the commercial vitality of all the commercial entities in this commercial grid, defined as (2).

\[
Y^g_t = \sum_{\tau \in g} y^\tau_t = \sum_{\tau \in g} (\sum_{i \in \tau, t} |\text{Review}_i| + \sum_{i \in \tau, t} |\text{Checkin}_i|) \quad (2)
\]

There is, in fact, no universally accepted definition of commercial vitality. However, it is generally believed that reviews and check-ins reflect visitors’ preferences. More reviews
## Majors finds of existing schemes for commercial vitality prediction

| Scheme                              | Key Finding                                                                 | Advantage                                                                 | Disadvantage                                                                 |
|-------------------------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Work of Wang et al. [28]            | A linear correlation is established between social contexts and commercial vitality. | It is based on simple assumptions with easy verification.               | The linearly predictive model fails to learn the features from the complex check-in data. |
| CNNE (He et al. [10])               | Adaptive multimodal features are explored for predicting commercial vitality. | It can yield better prediction than those based on single view data.          | The prediction precision is not satisfactory due to using only fully connected layers to integrate multi-tasks. |
| Work of Zhang et al. [36]           | CNN is introduced to predict urban congestion.                             | It captures the correlations between adjacent grids.                      | It includes the irrelevant relationships of remote entities and does not consider temporal contexts. |
| ST-3DNets (Guo et al. [9]), works of Ji et al. [14] and Tran et al. [27] | 3D CNN is introduced for prediction tasks.                                  | It captures both the spatial and temporal correlations.                    | It includes the irrelevant relationships of remote entities. |
| Works of Chen et al. [4] and Kong et al. [15] | LSTM is introduced to forecast urban housing price and urban power load. | It effectively captures the temporal feature.                             | It fails to reflect the spatial relationships and periodically temporal features. |
| Works of Huang et al. [12], Li et al. [17] and Shi et al. [23] | CNN and LSTM are integrated together for prediction.                     | It captures both the spatial and temporal correlations.                    | It may lead to gradient vanishing and exploding. |
and check-ins usually indicate a larger volume of visitors, resulting in a more popular and active corresponding district. Of course, the business performance of individual entities, such as revenues and profits, can also be used to indicate commercial vitality. However, they are commercial secrets with only limited access. Thus, in this research, the changes in commercial vitality is captured through reviews and check-in records from social apps, which adds a new perspective to the commercial vitality study.

3.2 Problem definition

**Commercial Vitality Prediction:** Given the commercial vitality in the commercial grid \( g(i, j) \) before a given time slot \( t \) (including \( t \)), the problem is to predict its commercial vitality at time slot, \( t + 1 \), as denoted by (3).

\[
P^g_{t+1} = \mathcal{F}(Y^g_0, Y^g_1, \ldots, Y^g_t)
\]

(3)

3.3 3D convolution

CNN is a type of deep neural network that mainly contains convolution and pooling layers within the hidden layers [24]. Here, convolution is a special linear operation, which can extract nonlinear features by combining them with activation functions. Furthermore, the size of the convolution kernel determines the range of feature extraction, known as the receptive field and can be interpreted as the sensory field of visual cortex cells [8]. When the convolution is carried out, the convolution kernel scans the input tensor, performs matrix multiplication and sums to stack the deviation amount on the scanned area in the receptive field [7]. Pooling is a form of nonlinear subsampling, which is mainly used to reduce the output tensor from the convolution layer and the number of parameters in the network [25]. Common pooling methods include mean pooling and maximum pooling.

When CNN is applied to image and video recognition, 2D and 3D CNNs are derived. The 2D CNN performs convolution, pooling and other operations on the input picture or video frame, to capture the plane visual pattern. Similarly, a deep cnn-lstm model uses 2D CNNs for particulate matter (PM2.5) [12] and a hybrid deep-learning algorithms uses 2D CNN to predict the distribution of dockless shared bicycles in the future [17]. However, due to the characteristics of plane, the spatio-temporal information between video frames cannot be modeled by 2D convolution in the temporal dimension. How to capture spatio-temporal information effectively without artificial annotation is a key problem in video convolution research. Ji et al. [14] utilize 3D CNN to recognize human action and Tran [27] et al. adopt 3D CNN to model spatial and temporal features for image classification tasks. They find that 3D CNNs could capture the spatial and temporal features more effectively than 2D CNNs. In other words, 2D CNNs can only extract spatial features from the neighborhood of the previous layer to obtain a feature graph, containing only spatial features. When a 2D CNN is applied to a spatio-temporal feature extraction of video, multiple consecutive frames are regarded as multiple color channels in the image. Thus, after the operation of 2D CNNs, multiple consecutive frames in the video are compressed into a plane’s feature graph, resulting in the loss of time information. By contrast, a 3D convolution uses a 3D convolution kernel to act on the cube generated by the superposition of multiple consecutive frames, and its feature map connects multiple successive frames in the previous layer, rendering it effectively in capturing spatio-temporal features. Compared with 2D CNN, both the input and output of 3D CNN are 3D matrices, enabling it to correctly extract and retain information relating to the time dimension. As shown in Fig. 1, different from 2D CNN, the size
of a convolutional kernel filter in 3D CNN is a $3 \times 3 \times 3$ cube. When performing the convolution, the local grids inside both the same and adjacent frame/time slots are simultaneously considered to compute the dependencies.

### 3.4 Residual network

Generally, deep neural networks can fit and capture high-dimension information better than shallow neural networks [11]. A good example is VGG [24], which greatly improves network performance by increasing network depth on the basis of AlexNet [16]. However, simply increasing the network depth does not guarantee satisfactory results in most cases. The reason is that the network depth increases rapidly with the complexity of the neural network model, resulting in gradient vanishing and exploding. Thanks to batch normalization and initial normalization, the deep learning model is effective in maintaining a good fitting performance, even at a deep network depth. However, when the depth of the network is further increased to capture complex, high-dimensional, nonlinear spatio-temporal relations, the neural network will degrade rapidly. To overcome this problem, a residual neural network is introduced into the model. Renowned for eliminating gradient vanishing and exploding, the residual neural network converts the learning task of each layer of the neural network into learning residual [11]. The experiment shows that the residual network is easier to optimize and can improve accuracy by increasing the appropriate depth. A residual neural network contains many residual units, which usually include two convolution layers.
and two batch normalization layers, as shown in Fig. 2. One residual unit is defined by (4):

\[ X_{l+1} = F_{res}(X^l) + X^l \]

where \( X^l \) and \( X^{l+1} \) are the input and output of the \( l^{th} \) residual unit, and \( F_{res} \) is a residual function.

### 4 Prediction of regional commercial vitality

This section introduces the detail of STCRNN. As is shown in Fig. 3, the framework is mainly composed of two elements, which are used to capture periodic and recent spatio-temporal relationships, respectively.

#### 4.1 Preprocessing

Commercial entities are spatially discrete and are not conducive to the capture of their spatial correlations. To eliminate this obstacle, firstly, the city map is divided, or rasterized,

![Residual unit diagram](image-url)
into a series of size-fixed grids according to the longitude and latitude coordinates. Then, the counts of reviews and check-ins relating to the commercial entities in each grid within a certain time slot are summed up, to denote the commercial vitality of the corresponding grid, as shown in Fig. 3. In this way, a commercial vitality map, is built with each grid being represented by a color (gray) value to indicate its commercial vitality. Here, the time span is set as one month. For each time slot \( t \in \Gamma \), the commercial vitality is rasterized and a set of grid vitality is obtained, represented as \( \xi_t = \{ P_{1}^{i,j}, P_{2}^{i,j}, \ldots, P_{m,n}^{i,j} \} \), in which \( i \) and \( j \) are the coordinates of grid centers. Figure 4 shows an example of rasterization.

According to the first law of geography [26] “near things are more related than distant things”, there is a close relationship between each grid of the commercial district and its surrounding grids. Therefore, a two-dimensional Gaussian blur is employed to simulate this relationship, by scattering the commercial vitality of each grid to its surrounding \( S \times S \) grids.

### 4.2 Extracting spatial correlations

In the previous mentioned spatio-temporal studies, the whole research area (i.e., the commercial vitality map) is fed into the CNN as an image. To focus more on correlations between the surrounding areas but eliminate the irrelevant relationships of remote areas, a local CNN is introduced to extract the spatial information of commercial vitality. The following steps demonstrate the process of image segmentation as shown in Fig. 4.

1. For each time slot \( t \), the commercial vitality map is placed in the first quadrant of a 2D coordinate system.
2. A grid, say \((i, j)\), and its surrounding grids are extracted as an \( S \times S \) image, starting from the origin of the coordinates, where \( S \) is an odd number and \((i, j)\) is at the center of the image, as shown in Fig. 5.
3. 

(i, j) is moved one grid to the right and the segmentation operation is repeated until the right edge of the last segmented image reaches the right border of the research area.

4. 

(i, j) is moved one grid up each time, until the upper edge of the segmented image reaches the upper border of the research area.

To date the map is segmented into M*N grids, and an image tensor set is obtained, denoted by \( \{Y_{ij}^{t} \mid 1 \leq i \leq m, 1 \leq j \leq n, 1 \leq t \leq T \} \), in which every image has the commercial vitality as a pixel value of each grid. The local CNN takes \( Y_{ij,k}^{t} \) as an input image to the \( k^{th} \) convolutional layer, which is defined by (5):

\[
Y_{ij,k+1}^{t} = f_1(Y_{ij,k}^{t} \ast W_k^t + b_k^t)
\]

where \( \ast \) denotes the operation of CNN and \( f_1 \) is an activation function. Here, \( Y_{ij,k}^{t} \) represents the vitality value of \( S \times S \) grids centered on grid \((i, j)\) within the \( t^{th} \) time slot (or frame) which is fed into the \( k^{th} \) CNN layer, and \( W_k^t \) and \( b_k^t \) denote the corresponding kernel weight and bias, respectively. Since the task is to predict the commercial vitality of the central grid within the \( S \times S \) grids, it is not necessary to apply any subsampling and pooling operations.

4.3 Extracting temporal correlations

The segmented image (grids), \( Y_{ij}^{t} \), at time slot \( t \), can be regarded as a video frame. Thus, a series of segmented images, with a same center grid at contiguous time slots constitutes a “video stream”. Here, a 3D convolution is applied to capturing temporal features from this “video stream”. Here, local CNN is applied to each 2D dimension of 3D convolution to capture spatial correlations, as the previous section illustrates. Considering the difference
between periodic and recent characteristics of commercial vitality, a periodic neural network is designed together with recent neural networks, to extract the temporal correlations.

When designing a periodic neural network, several 3D convolutions are applied to extracting periodic, spatio-temporal correlations. For each 3D convolution, the activation function is represented by (6):

\[ \tilde{Y}_{ij,k}^{t+1} = f_2(Y_{ij,k}^{t} \ast W_k^t + b_k^t) \]

where \( f_2 \) is an activation function, \( \ast \) denotes the operation of 3D convolution, \( \tilde{Y}_{ij,k}^{t+1} \) represents the output of convolution units, \( Y_{ij,k}^{t} \) is the input of the \( k^{th} \) 3D convolution layer and \( W_k^t \) and \( b_k^t \) are learnable parameter sets. Furthermore, a residual unit consisting of an identity function and a 3D convolution is given by (7):

\[ Y_{ij,k}^{t+1} = \tilde{Y}_{ij,k}^{t+1} + Y_{ij,k}^{t} \]

where \( \tilde{Y}_{ij,k}^{t+1} \) and \( Y_{ij,k}^{t} \) are the outputs of residual units in the \((k + 1)^{th}\) and \( k^{th} \) layer respectively. After the 3D convolution layers and residual units, the output tensor \( Y_{ij}^{t} \) is flattened to a feature vector \( v_{ij}^{t} \) for grid \((i, j)\) at time slot \( t \). Finally, a fully connected layer is applied to reduce the length of the spatio-temporal feature vector, \( v_{ij}^{t} \), which is defined by (8):

\[ \eta_{ij}^{t} = f_3(W_{FC^{3d}} v_{ij}^{t} + b_{FC^{3d}}) \]

where \( f_3 \) is an activation function, and \( W_{FC^{3d}} \) and \( b_{FC^{3d}} \) are learnable parameter sets. So far, a periodic, spatio-temporal vector \( \eta_{q,periodic}^{ij} \) is obtained. Thus, the periodic correlations in \( q_{periodic} \) time slots generate \( t \) vectors of \( \eta_{q,periodic}^{ij} \), which are then combined into a single tensor \( \eta_{q,periodic}^{ij} \). Here, \( q_{periodic} \) represents the number of periodic time slots.

When designing the recent neural network, it is assumed that recent correlation is more relevant to the prediction result. Based on this idea, the depth of recent neural networks, i.e., the number of 3D convolution layers, is increased. Since increasing layers leads to gradients vanishing, the residual network is introduced to extract recent information. As mentioned above, in each residual unit, two 3D convolution layers and two batch normalization layers are employed, as Fig. 3 shows. Following a flattened layer and a fully connected layer, a recent spatio-temporal tensor \( \eta_{q,recent}^{ij} \), is obtained, where \( q_{recent} \) represents the number of recent time slots.

In order to automatically assign different weights to recent correlation and periodic correlation, a LSTM [22] network is applied to the model. The two tensors \( \eta_{q,periodic}^{ij} \) and
\( \eta_{q, \text{recent}} \) are concatenated as (9).

\[
\theta_{ij}^t = \eta_{q, \text{periodic}} \oplus \eta_{q, \text{recent}}
\]   

(9)

Therefore, the tensor, \( \theta_{ij}^t \), contains \( q_{\text{recent}} + q_{\text{periodic}} \) periods of spatio-temporal information. \( \theta_{ij}^t \) is then fed into LSTM and an output tensor, \( \delta_{ij}^t \), is obtained, as (10) indicates.

\[
\delta_{ij}^t = \text{LSTM}(\theta_{ij}^t)
\]   

(10)

As shown in Fig. 6, there are three different logic gates controlling the flow of information in the hidden layer of LSTM. The input gate (\( \text{input}_k^t \)), controls the input of the memory unit, the forgetting gate (\( \text{forget}_k^t \)), controls the forgetting of the memory unit and the output gate (\( \text{output}_k^t \)), controls the output of the memory unit. Here, \( k \) denotes the \( k \)th memory unit in the hidden layer of LSTM. Assuming that \( h_k^{t-1} \) denotes the hidden layer output, \( c_k^{t-1} \) denotes the state of memory unit state and \( x_k^t \) denotes the coded value of \( \theta_{ij}^t \), the hidden layer output, \( o_k^t \) and the cell state, \( c_k^t \) of \( \theta_{ij}^t \) can be calculated using the following formulas.

Firstly, apply activation function, \( F_1 \), to control the information to be discarded from \( x_k^t \) and \( h_k^{t-1} \), which is given by (11):

\[
\text{forget}_k^t = F_1(W_{\text{forget}}^k \cdot [h_k^{t-1}, x_k^t] + b_{\text{forget}}^k)
\]   

(11)

where \( W_{\text{forget}}^k \) and \( b_{\text{forget}}^k \) are learnable parameters sets.

Then, update the input gate to control the information to be stored by (12):

\[
\text{input}_k^t = F_2(W_{\text{input}}^k \cdot [h_k^{t-1}, x_k^t] + b_{\text{input}}^k)
\]   

(12)

where \( F_2 \) is the sigmoid activation function, \( W_{\text{input}}^k \) and \( b_{\text{input}}^k \) are learnable parameter sets. The updated candidate value, \( g_k^t \), is calculated by (13).

\[
g_k^t = F_3(W_g^k \cdot [h_k^{t-1}, x_k^t] + b_g^k)
\]   

(13)

where \( F_3 \) is the tanh activation function, \( W_g^k \) and \( b_g^k \) are the learnable parameter sets. After calculating \( \text{input}_k^t \) and \( g_k^t \), the previous state, \( c_k^{t-1} \), can be updated to the new state, \( c_k^t \), by (14).

\[
c_k^t = \text{forget}_k^t \cdot c_k^{t-1} + \text{input}_k^t \cdot g_k^t
\]   

(14)

where \( \text{forget}_k^t \cdot c_k^{t-1} \) represents the information to be forgotten in the previous cell state, and \( \text{input}_k^t \cdot g_k^t \) represents the information to be added to the next memory cell state. Then, the output information can be calculated through the \( F_4 \) activation function, which is denoted by (15):

\[
\text{output}_k^t = F_4(W_{\text{output}}^k \cdot [h_k^t, x_k^t] + b_{\text{output}}^k)
\]   

(15)

Fig. 6 The structure of LSTM
where $F_4$ is the sigmoid activation function, $output^k_t$ represents the output gate in the memory unit, and $W^k_{output}$ and $b_{output}$ are learnable parameter sets. The next hidden layer output, $o^k_t$, and the next unit output, $h^k_t$, can be obtained by (16), where $F_5$ is the sigmoid activation function.

$$o^k_t(h^k_t) = output^k_t \ast F_5(c^k_t)$$

By connecting all $o^k_t$ corresponding to $\theta_{ij}^t$, it is obtained that $\delta_{ij}^t = o^1_t \oplus o^2_t \oplus ... \oplus o^k_t$, where $\oplus$ represents vector connection.

### 4.4 Prediction

For the final prediction, the goal is to obtain all grid values of commercial vitality for the future time slot, $t + 1$. Since the tensor, $\delta^{ij}$, already contains spatial and temporal correlations, a fully connected layer is applied to calculating the final prediction value, $\hat{P}^{ij}_{t+1}$, which is given by (17):

$$\hat{P}^{ij}_{t+1} = f(W^{FC}_t \delta^{ij}_t + b^{FC}_t)$$

where $W^{FC}_t$ and $b^{FC}_t$ are learnable parameter sets, $f$ is the activation function and $\delta^{ij}_t$ denotes the output of LSTM. The final result is normalized to $[0,1]$ and can be denormalized back to real commercial vitality values. According to the order that the local CNN segments images, the predicted values are calculated back to obtain the predicted grid value of commercial vitality.

### 4.5 Loss function

The model is trained by minimizing the value of loss function iteratively, which is defined as the Mean Absolute Error (MAE) between the real commercial vitality value and the predicted value, as (18) shows:

$$L_{\theta} = \frac{1}{m \ast n} \sum_{i=1,j=1}^{m,n} |P^{ij}_t - \hat{P}^{ij}_t|$$

where $\theta$ represents all learnable parameters in STCRNN, $P^{ij}_t$ denotes the real vitality value, $\hat{P}^{ij}_t$ denotes the predicted vitality value and $m \ast n$ represents the number of samples.

During each iteration of the training process, all the required parameters, $\theta$, will receive an update, known as epoch in deep learning. The condition according to which the model terminates training, occurs when the hyper-parameter epoch reaches the required number of times set. However, the determination of epoch often requires comprehensive consideration of training completion and overfitting. From Fig. 7, it can be seen that the optimum epoch value occurs, in theory, at the highest point of the test set accuracy. When the epoch falls below this ideal value, the accuracy of the training set and test set is still low, due to insufficient training. When the epoch increases above this ideal value, the model overfits the data of the training set, also leading to lower accuracy of the test set. Therefore, an early stopping strategy is adopted in the training process. In other words, the training process is stopped in advance, when the error of the test set is greater than the error of the previous iteration.
5 Experiments

This section presents the experimental settings as well as the experimental results which demonstrate the effectiveness of the proposed model.

5.1 Experimental settings

This section briefly introduces the experimental settings, including the datasets, parameter settings, evaluation metrics and comparison methods.

5.1.1 Datasets

The public dataset is downloaded from Yelp\(^1\) for the experiments, which includes more than 668,000 reviews and more than 160,000 check-in records of stores in various cities around the world from October 2004 to November 2018. As Las Vegas and Toronto have the largest number of reviews among all cities, they are selected as the experimental cities.

After investigation, the data from January 2013 to December 2017 are selected as the training dataset, and the data from 2018 as the testing dataset, since the reviews and check-ins before 2013 are inadequate. Table 2 provides the statistics of the dataset. In the experiment, a period of one month is adopted as the time span. Besides, the same month for the previous three years and the previous 11 months are used to predict the commercial vitality for the next month. Besides, the vitality values less than 10 are filtered out, which is common practice in a practical application [32].

5.1.2 Parameters settings

STCRNN is implemented by Keras, a fast neural network API running on top of TensorFlow. The experiments run on a cluster with four NVIDIA 1080Ti GPUs. During experiments, the min-max normalization is applied on training datasets, to normalize input

---

\(^1\)https://www.yelp.com/dataset
values to [0,1]. After the prediction, the min-max normalization is reversed to recover commercial vitality values.

In the experiment, the size of the surrounding grids is set to 7 * 7, which corresponds to an actual 700m * 700m rectangular area. For the periodic element of STCRNN, twelve 3D convolution layers with 32 filters are employed, of which the size is set to 3 * 3 * 3. For the recent element, the number of residual units is set to 12. In addition, 32 filters with a size of 3 * 3 * 3 are applied to the 3D convolution layers of each residual unit. In all the convolution layers, ReLU is adopted as the activation functions. For LSTM, the dimension of the hidden state vector is set to 512. In the fully connected layer, Sigmoid is adopted as the activation function. With regard to the other parameters, the batch size is set to 64, and the learning rate is set to 0.001. The best combination of above parameters is obtained by a greedy searching approach in a possible range. Finally, Adam is applied as the optimizer.

5.1.3 Evaluation metric

Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) are used as accuracy indicators of the model, as defined in (19) and (20).

\[
MSE = \frac{1}{m \times n} \sum_{i=1,j=1}^{m,n} (P_{ij}^t - \hat{P}_{ij}^t)^2
\]

\[
MAPE = \frac{1}{m \times n} \sum_{i=1,j=1}^{m,n} \left| \frac{P_{ij}^t - \hat{P}_{ij}^t}{P_{ij}^t} \right| \times 100
\]

where \(P_{ij}^t\) is the observed vitality value, \(\hat{P}_{ij}^t\) is the predicted value for the grid \((i, j)\) at time slot \(t\) and \(m \times n\) represents the number of samples.

5.1.4 Comparison methods

The presented model is compared with five other models. In addition to the proposed STCRNN, the multiple experiments of its variants, i.e., STCRNN without residual network and STCRNN without periodic information, are also conducted.
- Historical Average (HA): a traditional prediction method, in which the predicted value is the average value of previous historical commercial vitality values.
- Autoregressive Integrated Moving Average (ARIMA) [1]: a traditional time series analysis method, which does not consider the variations of other relevant, random variables.
- Convolution LSTM (ConvLSTM) [23]: by extending the fully connected LSTM (FC-LSTM) to include convolutional structures in both the input-to-state and state-to-state transitions, ConvLSTM can be used to build an end-to-end trainable model for the now-casting problem.
- ST-3DNets [9]: a recent traffic-forecasting method, which simulates the change of adjacent crowd flow through the deep convolution residual network.
- XGBoost [2]: a gradient-boosting tree model, which has been very popular recently after many winning teams participating in competitions used it.
- STCRNN without residual network: a variant version of STCRNN, which does not contain the residual network.
- STCRNN without periodic dimension: a variant version of STCRNN, which does not contain the neural network structures to extract periodic variations.

All the parameters for the baselines are tuned to achieve their best performance on the testing dataset. Table 3 presents the details of modeling with XGBoost in the experiment. As for ConvLSTM, 8, 16 and 32 convolution layers with size 3 \( \times \) 3 filters are tried. The result shows it performs best when the depth is set to 16, the length of filters is set to 1 and the number of LSTM layer is set to 1. Similarly, as for ST-3DNets, 8, 16, 32 and 64 convolution layers are tried with filters of size 3 \( \times \) 3 \( \times \) 3. It is found that ST-3DNets performs best when the depth is set to 32 and the length of filters is set to 32.

### 5.2 Experiment results

This section first demonstrates the experimental results of STCRNN, compared with the other models. Subsequently, it shows how the parameters (i.e., size of surrounding grids, number of recent months and number of residual units) influence the model performance. Finally, it provides the visualization results of Las Vegas and Toronto using STCRNN for the year 2018.

#### 5.2.1 Model performance

Table 4 shows the average results of ten rounds of experiment, which indicates that STCRNN achieves the lowest MSE (i.e., 26.25) and the lowest MAPE (i.e., 16.20) across all models. Compared to the second-best model (i.e., ST-3DNets), STCRNN improves the

| Attribution      | value |
|------------------|-------|
| learning rate    | 0.09  |
| n-estimators     | 90    |
| max depth        | 9     |
| min child weight | 5     |
| subsample        | 0.85  |
| colsample bytree | 0.8   |

...
Table 4 Comparison of different models

| Model                                | MSE  | MAPE |
|--------------------------------------|------|------|
| HA                                   | 53.79| 22.98|
| ARIMA                                | 61.76| 21.92|
| ConvLSTM                             | 29.95| 18.31|
| XGBoost                              | 28.68| 22.15|
| ST-3DNets                            | 26.53| 16.45|
| STCRNN                               | 26.25| 16.20|
| STCRNN without periodic dimension    | 27.50| 17.47|
| STCRNN without residual units        | 567.70| 100.00|

performance by 1.1% in MSE and 1.5% in MAPE, respectively. It is worth noting that although ST-3DNets also captures the recent and periodic variations, its performance is still worse than that of the proposed model. There are two possible reasons for it. The first one is that ST-3DNets fails to aggregate two kinds of temporal dependencies due to the lack of LSTM. Another reason is that the proposed model utilizes the local CNN, and only feeds it the vitality of a commercial district within a certain range of neighboring grids, i.e., 7×7 grids, rather than computing the whole commercial map at one time, so as to remove irrelevant correlations between distant commercial districts. Although the improvement over ST-3DNets is not so significant, it is still very remarkable as STCRNN is the first deep-learning-based approach to predict the vitality of commercial districts based on publicly available online reviews and the check-in records of commercial entities. Besides, XGBoost is an efficient and scalable implementation of gradient boosting framework [3]. Nevertheless, it is not capable of processing high-dimensional data with strong spatial-temporal dependencies, which leads to unsatisfactory performance in the prediction task of commercial vitality. As for two traditional models, both perform much worse than the proposed one, among which HA simply computes the historical average value of grids in the corresponding time slot/frame, and ARIMA requires data to be stationary in the sequence, leading to the very poor performance on the complex check-in data. Generally speaking, compared with the traditional ones such as HA and ARIMA, machine learning and deep learning models obviously demonstrate better performances.

Additional, part of the model structure is removed for the ablation experiment. The result, indicated in Table 4, shows that STCRNN without periodic dimension, degrades by 4.8% in MSE and by 7.8% in MAPE, respectively. It demonstrates the importance to learn the periodic features for improving the prediction performance. When the residual neural network is not integrated into the model, the prediction sharply fails, as shown also in Table 4, due to multiple convolution layers. This result indicates the significant importance of the residual neural network in preventing gradient vanishing and exploding.

In order to judge the stability and significance of the model’s result, the F-test and T-test are performed. As Table 5 shows, the p-value of F-test is significantly greater than 0.05, indicating that the predicted results are consistent with the correct label or the robustness.

Table 5 F-test and T-test of STCRNN

|                | F test | T test |
|----------------|--------|--------|
| p-value        | 0.3694 | 0.0017 |
of the test is quite reliable. Meanwhile, the p-value of T-test is significantly less than 0.05, indicating that it is statistically significant.

5.2.2 Parameter influence

The second experiment explores how the number of residual units influences the accuracy of the model. As Fig. 8a shows, when the number of residual units ranges between 11 and 13, both MSE and MAPE are relatively stable. Otherwise, MSE rises significantly due to the lack of deep mining of commercial vitality. Because MAPE better reflects the overall prediction accuracy of the model, 12 residual units are applied in the model.

In addition, since the size of the surrounding grids determines the input image size of the local CNN, the best grid size for the model is also estimated. As shown in Fig. 8b, when the size selected is 700m * 700m, both MSE and MAPE obtain their optimal values. MSE and MAPE increase slightly when the size decreases to 500m * 500m and a lower number of surrounding relationships are taken into account. When the size is increased to 1.3km*1.3km or 1.5km * 1.5km, MSE and MAPE rise significantly. The reasons may be as follows: 1) several unrelated locations are included in the CNN and hence reduce the precision of the model when the size becomes enlarged. 2) a larger grid results in a decline of the sample size. If the grid continues to expand and eventually covers the whole research area (the whole area becomes a single grid or image which is fed into the CNN),
the performance will be degraded quickly and will become unacceptable, according to the experiment.

After exploring the important influence of periodic dimension on the experimental results in the first experiment, it is also attempted to establish the influence of recent data on the results. Figure 8c shows that prediction accuracy is at its highest level when the data of the previous 11 months are included as the input of the model. When the number of recent months is less than five, the results fluctuate significantly, due to the fact that only using several preceding months would lead to overfitting. However, at certain points, the results are surprisingly satisfactory.

With regard to the other parameters, extensive experiments are implemented to determine their best values while preventing overfitting, based on the validation dataset. The operations of batchnormalization, i.e., the regularization item, are added after each of the 3D convolution layers to prevent overfitting. Furthermore, the model performance is tested by trying the different values of Dropout from 0 to 1 on each layer to find its best value. Finally, it is even considered to increase the number of validation samples. The experimental result shows that the model, based on the current validation set, has already obtained very similar prediction results to the real situation.

5.2.3 Visualization results

Figures 9 and 10 demonstrate the visualization results of the distribution of commercial vitality. For better illustration, the most inactive parts have been masked. As they indicate, Las Vegas is more commercially active than Toronto, since it is famously known as the ‘sin city’, with more commercial entities, such as casinos, restaurants, malls and night clubs. Meanwhile, as Fig. 9 shows, both the real and predicted results for Las Vegas, indicate that this location is more commercially active in summer than in other seasons, which is consistent with the real situation. However, the difference becomes slightly more obvious in the case of the real Las Vegas than the predicted model. As for Toronto, the model predicts the real hot business centers with almost the same shapes in terms of commercial vitality, as Fig. 10 shows.

6 Discussion

The novelty of STCRNN, compared to the state of the art, mainly includes the following two aspects. As for the spatial features, different from the traditional CNN that processes pixels all over the map, STCRNN employs the local CNN and only extracts the dependencies of surrounding commercial district, so as to remove the irrelevant influence from distant commercial districts. As for the temporal features, few existing models on commercial vitality prediction exploit the periodic features. By contrast, STCRNN learns not only the recent but also periodic variations to improve the prediction effect. More specifically, STCRNN utilizes the 3D convolution to study the correlations between surrounding grids in recent and periodic time slots. Furthermore, different from the models like ST-3DNets, which aggregates the recent and periodic features together in a weighted way, STCRNN utilizes LSTM to retain the information from the recent and periodic features, thus improving the prediction accuracy.

In fact, the effectiveness of the proposed model can be attributed to its different structures. The 3D CNN has been verified to be capable of learning the correlations of
Fig. 9 Visualization results of real and predicted vitality of Las Vegas in 2018
Fig. 10 Visualization results of real and predicted vitality of Toronto in 2018
spatial-temporal representation in some complicated scenarios such as traffic volume prediction [9] and video analysis [14, 27]. Therefore, the 3D CNN technique is utilized to study the spatial-temporal characteristics from the check-in dataset. Different from traditional approaches, the proposed model only considers the certain range of commercial grids, i.e., the neighboring grids around the target grid, to reduce the irrelevant correlations. Additionally, a 3D CNN channel is employed to learn the periodicity of the commercial vitality over time. Thus, the model is enhanced to capture the variation of commercial vitality more precisely. Finally, two temporal features are integrated by using LSTM, in which each value of the output neural unit in the last fully connected layer is taken as a dimension of the input vector in LSTM. In this way, the proposed model eventually wins others.

The proposed model consists of three modules, among which the computational complexity of 3D CNNs module is significantly larger than the fully connected layers and the LSTM module. Therefore, the computational complexity of the whole model can be determined by that of the 3D CNNs module, i.e., \( O(L^2 \times \sum_{i=1}^{D} M_i^3 K_i^3 C_i^2) \). Here, \( L \) denotes the map size, \( D \) denotes the number of 3D CNN layers, and \( M_i \), \( K_i \) and \( C_i \) denote the output feature map size, the filter kernel size, and the number of filter kernels in the \( i^{th} \) layer respectively. For the application here, there are twelve 3D CNN layers, or \( D=12 \). Meanwhile, \( M_i \) equals 7, \( K_i \) equals 3 and \( C_i \) equals 32, for all layers. Thus, for a fixed setting of the model, its computational complexity is \( O(L^2) \). As a matter of fact, in the experiment, it takes averagely 92.59 and 102.74 mins in total 10 epochs for 10 times training on the Toronto and Las Vegas city datasets on four NVIDIA 1080Ti GPUs, respectively. The computational complexity and running performance are acceptable in real situations. Of course, the model can be accelerated by leveraging the power of cloud computing, if huge amount of check-ins and reviews needs to be handled. Nevertheless, it is imperative to guarantee the data security when they are transmitted through Internet [21].

Meanwhile, there are also some limitations in this work. One may come from the scale of experiment. Due to the limited data, the model is only evaluated on Las Vegas and Toronto. Nevertheless, the model is believed to be also effectively applied to other cities, because it has been successfully evaluated on the randomly selected sample data. Another limitation is that it does not consider the accidental events which may seriously affect the development of commercial district in a drastic way, such as the COVID-19 outbreak. However, to deal with such accidents is a well-recognized difficulty for all prediction tasks.

In real situation, the regional commercial vitality usually evolves as time goes on. To accurately predict the future commercial vitality helps many tasks, such as investment decision and urban planning. This would be one of cores of so-called city brain or smart city. More specifically, if one certain region is predicted to be booming, it is suggested to be worth opening more businesses there. Correspondingly, more facilities are required to be equipped, such as gas and water supplies, bus stops, clinics and even schools. On the other hand, the probabilities of success or failure of individual business entity can also be predicted based on reviews and check-in records collected from mobile apps. Obviously, such result makes the business operation less reactive but more proactive.

### 7 Conclusion

In this paper, a novel neural network, called STCRNN is proposed, which employs online reviews and check-in records to predict commercial vitality by month. Specifically,
STCRNN includes a spatial dimension that employs local CNN to capture the spatial relationship of surrounding commercial districts and a temporal dimension that applies 3D convolutions to deal with the temporal characteristics of commercial vitality. Considering the obvious periodic and recent patterns of commercial vitality, STCRNN handles periodic and recent temporal dimensions simultaneously and applies LSTM to combine both. In addition, to prevent the gradients vanishing and exploding, caused by an increase in convolution layers, the residual network is applied in STCRNN. Experiments on public Yelp datasets from 2013 to 2018 demonstrate that STCRNN outperforms other methods.

As for the future works, it is necessary to add more related information regarding commercial vitality, such as the traffic condition, as an input of the prediction, rather than reviews and check-in records only. Furthermore, the sentimental analysis of reviews from LBSNs will be incorporated to make the prediction more precise. Last but not least, the business trend (success or failure) can be investigated for the individual entity, based on the predicted commercial vitality of its surrounding business district.

Declarations

Financial Interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Box GE, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 65(332):1509–1526
2. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
3. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H (2015) Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4)
4. Chen X, Wei L, Xu J (2017) House price prediction using LSTM. arXiv:1709.08432
5. Clarke KC, Gaydos LJ (1998) Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. Int J Geogr Inf Sci 12(7):699–714
6. Du Z, Wu S, Zhang F, Liu R, Zhou Y (2018) Extending geographically and temporally weighted regression to account for both spatiotemporal heterogeneity and seasonal variations in coastal seas. Ecological Info 43:185–199
7. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, MIT press, pp 201–202
8. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Cai JF, Chen T (2018) Recent advances in convolutional neural networks. Pattern Recogn 77:354–377
9. Guo S, Lin Y, Li S, Chen Z, Wan H (2019) Deep Spatial–Temporal 3D convolutional neural networks for traffic data forecasting. IEEE Trans Intell Transp Syst 20(10):3913–3926
10. He Z, Yang S (2018) Multi-view commercial hotness prediction using context-aware neural network ensemble. proceedings of the ACM on interactive, mobile. Wearable and Ubiquitous Technologies 2(4):1–19
11. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recognit, pp 770–778
12. Huang CJ, Kuo PH (2018) A deep cnn-lstm model for particulate matter (PM2.5) forecasting in smart cities. Sensors 18(7):2220
13. Huang J, Obracht-Prondzynska H, Kamrowska-Zaluska D, Sun Y, Li L (2021) The image of the City on social media: A comparative study using “Big Data” and “Small Data” methods in the Tri-City Region in Poland. Landsc Urban Plan 206:103977
14. Ji S, Xu W, Yang M, Yu K (2012) 3D Convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
15. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2017) Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 10(1):841–851
16. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp 1097–1105
17. Li Y, Shuai B (2020) Origin and destination forecasting on dockless shared bicycle in a hybrid deep-learning algorithms. Multimed Tools Appl 79(7-8):5269–5280
18. Liu Y, Zheng Y, Liang Y, Liu S, Rosenblum DS (2016) Urban water quality prediction based on multi-task multi-view learning. Proceedings of the 25th international joint conference on artificial intelligence, pp 2576–2582
19. Mandhachitara R, Shannon R (2016) The formation and sustainability of same product retail store clusters in a modern mega city. J Econ Soc Geogr 107(5):567–581
20. Namasudra S (2019) An improved attribute-based encryption technique towards the data security in cloud computing. Concurr Comput: Pract Experience 31(3):e4364
21. Namasudra S (2020) Fast and secure data accessing by using DNA computing for the cloud environment. IEEE transactions on services computing
22. Schmidhuber J, Hochreiter S (1997) Long short-term memory. Neural Comput 9(8):1735–1780
23. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS), pp 802–810
24. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
25. Sun M, Song Z, Jiang X, Pan J, Pang Y (2017) Learning pooling for convolutional neural network. Neurocomputing 224:96–104
26. Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(sup1):234–240
27. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE international conference on computer vision, pp 4489–4497
28. Wang L, Gopal R, Shankar R, Pancras J (2015) On the brink: Predicting business failure with mobile location-based checkins. Decis Support Syst 76:3–13
29. Wrenn DH, Sam AG (2014) Geographically and temporally weighted likelihood regression: Exploring the spatiotemporal determinants of land use change. Reg Sci Urban Econ 44:60–74
30. Xu Y, Chen D, Zhang X, Tu W, Chen Y, Shen Y, Ratti C (2019) Unravel the landscape and pulses of cycling activities from a dockless bike-sharing system. Computers. Environ Urban Syst 75:184–203
31. Yang S, Wang M, Wang W, Sun Y, Gao J, Zhang W, Zhang J (2017) Predicting commercial activeness over urban big data, Proceedings of the ACM on Interactive, Mobile. Wearable and Ubiquitous Tech 1(3):1–20
32. Yao H, Wu F, Ke J, Tang XF, Jia Y, Lu SY, Gong PH, Ye JP, Li ZH (2018) Deep multi-view spatiotemporal network for taxi demand prediction. Thirty-second AAAI conference on artificial intelligence, pp 2588–2595
33. Yu D, Wang X, Sun X (2020) Prediction of regional commercial activeness and entity condition based on online reviews. Int J Softw Eng Knowl Eng 30(11&12):1689–1705
34. Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 186–194
35. Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175
36. Zhang J, Zheng Y, Qi D (2017) Deep spatio-temporal residual networks for citywide crowd flows prediction. Thirty-first AAAI conference on artificial intelligence, pp 1655–1661

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Affiliations

Dongjin Yu¹ · Xinfeng Wang¹ · Ping Liang¹ · Xiaoxiao Sun¹

Xinfeng Wang
kaysen@hdu.edu.cn

Ping Liang
liangpingprivate@gmail.com

Xiaoxiao Sun
sunxiaoxiao@hdu.edu.cn

¹ School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China