Yu *et al. Journal of Big Data* (2022) 9:83
https://doi.org/10.1186/s40537-022-00637-9

Journal of Big Data

RESEARCH Open Access

Traffic flow prediction based on depthwise separable convolution fusion network

Yue Yu, Wei Sun, Jianhua Liu* and Changfan Zhang

*Correspondence:
jhliu@hut.edu.cn

School of Railway Transportation, Hunan University of Technology, Zhuzhou 412001, China

Abstract

Traffic flow prediction is an important part of an intelligent transportation system to alleviate congestion. In practice, most small and medium-sized activities are not given priority in transport planning, yet these activities often bring about a surge in demand for public transport. It is recognized that such patterns are inevitably more difficult to predict than those associated with day-to-day mobility, and that forecasting models built using traffic data alone are not comprehensive enough. Aiming at this problem, a depthwise separable convolutional fusion forecast network (FFN) was proposed by focusing on the impact of event information on traffic flow demand. FFN fused heterogeneous data to model traffic data, weather information, and event information extracted from the Internet. The depthwise separable one-dimensional convolution was used to encode the textual information describing the event layer by layer, and local one-dimensional sequence segments (ie subsequences) were extracted from the sequence to retain rich local semantic features. In the modeling process, the interaction of heterogeneous data was established, that is, the temporal and other data were used to drive the textual information representation in the encoding process to capture better relevant textual representations. Finally, information from different sources and formats was fused to obtain a joint feature representation tensor that predicts the traffic demand in the next day's event area. The experimental results show that the average absolute error of the fusion prediction network is reduced by 26.5%, the root mean square error is reduced by 11.6%, and the judgment coefficient is increased by 26.4% compared with the prediction network that only considers the traffic data.

Keywords: Deep learning, Taxi demand prediction, Event area, Textual data, Heterogeneous data fusion

Introduction

Traffic flow forecasting is an important technology for alleviating urban traffic congestion, and one of the great challenges is to adequately adapt the supply of public transport to demand. Hao et al. [1] took the fitness function as the measurement standard and used the differential evolution algorithm to optimize the parameters of the radial basis function to obtain the optimal short-term traffic flow prediction value, thereby improving the short-term traffic flow prediction accuracy. Fang et al. [2] introduced an attention mechanism that assigns different weights to different inputs in a long short-term memory (LSTM) network, helping the network model to

make accurate predictions. Rajalakshmi et al. [3] reduced the prediction error rate by mixing convolutional neural network (CNN) and LSTM models. However, none of them considered the impact of regional events on the forecast. In traffic planning, traffic demand forecasting for special events is a well-known challenge. Transportation systems are usually designed according to habitual demands, and only very large events (eg concerts, major festivals, olympic games, world cups) receive special attention, and thus face great challenges in interpreting non-habitual transportation demand scenarios. In this scenario, in order to explain the reasons for non-habitual traffic demands, certain background knowledge is needed to discern the explanation, so this paper fully considers the impact of events in the predictive modeling process.

Previous studies have shown [4–6] that the information contained in real-time online resources such as announcement websites and social networks does have practical value for urban traffic demand forecast modeling. The abundance of information on public events on the Internet helps explain observed real-world phenomena, such as non-habitual overcrowding scenarios. With the advent of the era of big data, the Internet has become a valuable data source for traffic flow modeling. For example, predicting customer churn behavior through twitter [7], using social media data to identify the distribution of relief material needs [8], and studying the impact of multimodal information and social network information on travelers' commute mode choices [9]. People can understand various event information in real time through the Internet, not only know what is happening now, but also know what may happen in the future. But getting the model to understand event information and predict traffic flow demand well is a challenging problem.

Over the past few decades, deep learning has achieved success in multimodal classification or clustering tasks such as video classification [10, 11], event detection [12, 13], sentiment analysis [14, 15] and visual question answering [16] good effect. Successes in these areas demonstrate that deep learning can handle heterogeneous data well. Of course, there is no exception in the field of traffic flow forecasting. Xiao et al. [17] fused various heterogeneous information such as historical traffic flow and location semantics through gating layers and hierarchical adaptive graph convolutional networks to learn spatiotemporal interactions and traffic correlations in different levels of spatial dimensions. Yu et al. [18]. proposed a cross-attention fusion spatiotemporal multi-graph convolutional network model to fuse temporal and spatial features separately to reduce the prediction error of traffic flow prediction. Rodrigues et al. [19] proposed a fully connected layer deep learning (DL-FC) network for event area taxi demand prediction. Using the Internet as a contextual information resource for special events, the proposed deep learning architecture can significantly improve the quality of predictions.

Existing traffic flow prediction methods mainly focus on using the short-term correlation of recent observation patterns [2, 20, 21] to capture the periodic movement trend related to habit / routine behavior. However, when trying to model time series, we often ignore the valuable text information in the form of unstructured data. To solve this problem, this paper considered the impact of unstructured text data describing events on regional traffic flow demand, and proposed a deep separable convolution FFN. Compared with DL-FC network, this network has fewer learning parameters, higher

efficiency of information representation and better prediction effect. The contributions of this paper are as follows:

1. The depthwise separable one-dimensional convolution (SConv1D) was employed to learn multi-word patterns describing event text. First, the word embedding dimensions were separated and an independent spatial convolution is performed on each dimension. Then, point-by-point convolution was performed on the spatially convoluted tensor to mix the information in the embedding dimension. This effectively alleviated the problem of too many convolutional layer parameters due to too long word sequences and too large word embedding dimensions.
2. An interaction mechanism between unstructured text data and structured temporal data was established. We explored the use of temporal and other structured data to drive the text feature representation in the early, middle and late stages respectively, and obtained the "reset" text abstract feature representation. Experimental data show that the text representation in this interactive mode is more efficient.
3. Detailed incremental experiments were designed to quantify the impact of different components on fusion prediction results. At the same time, the contribution of different modal information sources to the model was evaluated. The importance of considering the textual information describing the event in the taxi demand prediction problem in the event area was highlighted.

The rest of this article is organized as follows. In the next section, the relevant literature for this work is reviewed. In "Proposed methodology" section, the neural network architecture for fusing structured data and unstructured text data is introduced. "Experiments" section describes the dataset used for the experiments and discusses the experimental results. The paper concludes with a conclusion ("Conclusion" section).

Related work

Urban flows and special events

Urban mobility demand forecasting has always been a challenging problem, and typical methods can only understand habitual behaviors and cyclical trends. However, our city is "dynamic", especially when some large-scale public events happen [22], which often leads to a surge in the demand for local taxis. The largest proportion of special activities is generally a variety of small and medium-impact activities. Their impact on mobility is difficult to measure, especially when multiple activities occur simultaneously, which undoubtedly poses a huge challenge to the task of traffic flow prediction.

From traffic planning's point of view, Chen et al. [23] introduced artificial neural network to predict traffic flow in different periods. Kuppam and Chang et al. [24, 25] proposed a conventional 4-step model to forecast special events' demands. Their model is highly dependent on manual survey data and does not include explicitly the event features that may impact the flows. Noursalehi et al. [26] proposed a single-variable and multi-variable state-space model, which is capable of simulating the impacts of external sources (such as soccer matches), and predicted in real-time the short-term traveler arrival volume. Yao et al. [27] proposed an exploratory method of digging twitter messages to understand the impact of people's activity patterns in prior evening/mid-night

on the flow on the next day morning. However, most of these works do not take into account the modeling of text information describing special events.

In reality, most of the information describing events is presented in the form of unstructured text. Due to the unstructured attributes of the text, the text cannot directly participate in the modeling of the character set whose variable type is required to be numerically related in data mining, but the text can be used as an important variable after processing. This paper integrates the text information describing events into the time series modeling, models the cross modal heterogeneous information, and fully considers the impact of special events on taxi demand forecasting.

Multi-modal data fusion

Multimodal data fusion [28] refers to various forms of combination of two or more modal data. For each source or form of information, it can be called a modality. Different modal data have different levels of knowledge expressivity to different degrees. Therefore, researchers have begun to focus on how to fuse data from multiple fields to achieve the complementarity of multiple heterogeneous information.

In view of the universality of multi-modal data fusion, there are many solutions for multi-modal data fusion in different fields. For instance, in a task of sentiment analysis, Zadeh et al. [29] proposed to fuse the voice, video, and audio modals into a tensor, which consists of the products of the specific feature vectors from all models, thereby exploiting the inner and intra sentimental dynamics. Fukui et al. [30] proposed to fuse the lingual and image modals using the multi-modal compact bi-linearity. They approximate by randomly projecting the images and texts onto space of higher dimensions and conduct effective convolution of these two vectors through elemental product in the FFT space. Liu et al. [31] proposed a multi-modal low-rank fusion method, which performs low-rank matrix decomposition on the weights and uses the low-rank tensors in the multi-modal fusion so as to improve fusion efficiency. Wu et al. [32] proposed multi-modal cyclic fusion method, which transforms the feature vectors into a cyclic matrix and fully utilize the interactions between the textual feature elements and visual feature elements. However, the abovementioned work contains comparatively few investigations on the fusion of time series data and textual information.

From the perspective of multi-modal data representation, deep multi-modal representation learning methods are generally divided into three frameworks [33]: joint representation, coordinated representation, and encoder-decoder. This paper adopts a joint representation strategy to input text data and temporal etc. data into different parts of the model at the same time, and uses the dependency between features to integrate different types of heterogeneous features to improve prediction performance. One of the advantages of joint representation is that it can easily integrate multiple modes, without explicitly coordinating the modes, and has strong versatility.

Proposed methodology

Textual representation based on SConv1D

The text is associated with the vector by using pre trained word embedding, that is, the embedded vector is loaded from the pre calculated embedded space Commonly used precomputed word embedding databases were word2vec [34], GloVe [35], etc. This paper

used GloVe embeddings. The prepared GloVe matrix was loaded into the embedding layer in Keras to obtain 3D word embedding vectors. The length of each sample sequence of this vector was 350 (after padding or clipping), and the dimension of each word embedding was 300, that is, the embedding layer returned a three-dimensional floating-point tensor with shape (samples, sequence_length = 350, embedding_dim = 300).

The sequence length of the above word embedding vector was very long and the word dimension was very high, which brought a large number of learning parameters to the text representation learning using neural network. In order to solve this problem, this paper did not use the traditional one-dimensional convolution method to learn text information, but used SConv1D [36] to extract text information.

Compared with ordinary convolutional layers, SConv1D can make the model lighter (fewer learning parameters) and faster (fewer floating-point operations). This layer carried out deep convolution and point-by-point convolution on the word embedding vector respectively. Firstly, the embedded dimensions (equivalent to "depth" or "channel" in CNN) are separated, and one-dimensional depth convolution is carried out on each embedded dimension, so as to learn multiple groups of local features between words. Secondly, each separated embedded dimension is spliced and then convoluted point by point to learn the inter channel mode, as shown in Fig. 1a.

The schematic diagram of the network structure is shown in Fig. 1b. After the embedding layer, the word embedding dimension is compressed. One-dimensional convolution with 30 convolution kernels of 1 was used to reduce the embedded dimension size. This was to reduce the learning parameters of SConv1D in the next recognition sequence.

Compressed text tensors passed through three SConv1D blocks in sequence. The internal details of the SConv1D block are shown in Fig. 1c. Each SConv1D contained a series of 1D convolutional filters that captured variable-length patterns in input text sequences of arbitrary size. Each one-dimensional convolutional layer used the tanh activation function instead of the relu activation function, in order not to lose information less than zero. A max pooling layer (Maxpooling) and a random dropout layer (Dropout) were also added after each layer. By subsampling the feature map and taking the maximum value as the feature corresponding to this particular filter. The size of Maxpooling was the same as the size of the previous 1D convolution filter kernel. Dropout dropped the feature representation of the previous layer with a 50% probability to reduce overfitting.

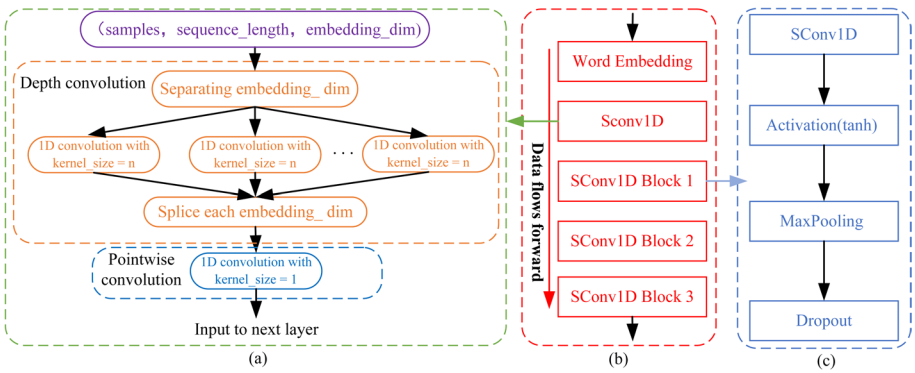

Fig. 1 **a** Details of the SConv1D; **b** Schematic diagram and network structure; **c** Details of the SConv1D block

Temporal guided textual representation

Unstructured text data often contains contextual interpretations of many patterns in traffic data. This makes it possible to establish a relationship between unstructured text data and structured temporal data. To better capture the salient underlying abstract feature representations of text in the taxi flow prediction task, temporal guided textual representation (TGTR) was proposed. TGTR guided the model to obtain important textual information for the current task as needed. A subset of text feature vectors at the corresponding stage was selected through structured temporal and other data, and secondary features were selectively filtered out, as shown in Fig. 2.

In the figure, $X \in R^d$ is the comprehensive information tensor obtained from the concatenation of time series information, weather information, and additional event information; d is the dimension after the concatenation of the features; $Y \in R^{m \times d'}$ is the textual tensor; m is the length of word sequence; d' is the dimension of word embedding, $d < d'$;$\odot$ denotes dot product; $\otimes$ denotes elemental multiplication; $\oplus$ denotes elemental addition.

Specifically, X first went through the batch normalization (BN) layer, the FC layer, and the BN layer in sequence. Then, did a dot product with Y to get α. BN used the mean and standard deviation of the mini-batch to adjust the network intermediate output, improving the stability of the intermediate layer output while minimizing overfitting. $\mathbb{W}$ in FC was a transformation matrix, which was also a linear correlation matrix. Secondly, sigmoid activation was performed on the result α of the dot product operation to obtain the class probability score β of each word. Each element in β represented a class probability that provided information conditioned on α, and words with outstanding contributions were given a larger class probability, and vice versa, a smaller class probability. Then, β went through the Permute layer, and the RepeatVector layer performed dimension transformation to obtain a tensor of the same dimension as Y, and multiplied it element-wise with Y. Finally, the residual connection obtained the "reset" textual abstract feature representation.

Taxi flow fusion forecast network

A schematic diagram of the entire network structure is shown in Fig. 3. It can be seen intuitively from the figure that the entire prediction network is a dual-input single-output type. The output is the predicted result under this input. Inputs include unstructured

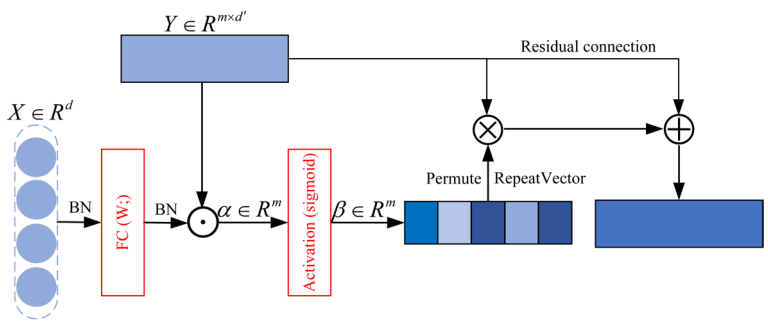

Fig. 2 Working mechanism of the TGTR

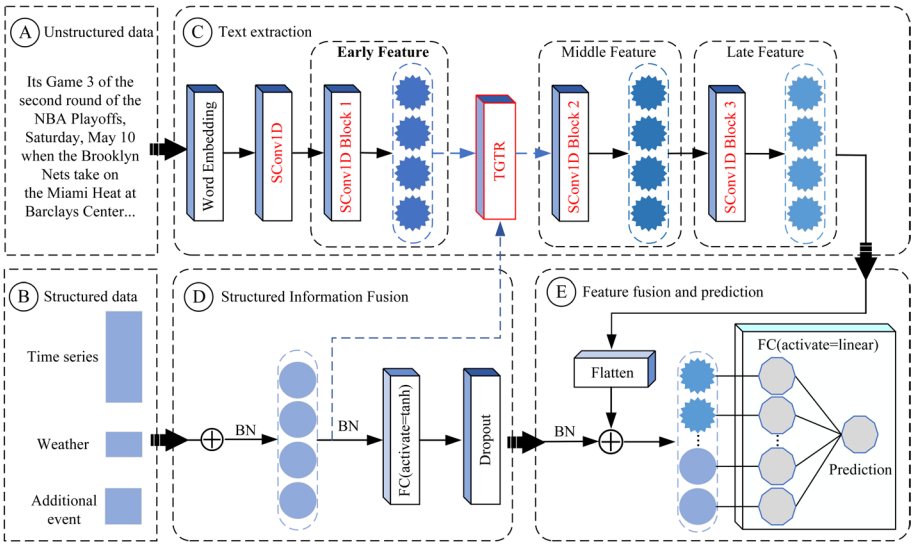

Fig. 3 FFN (SConv1D + earlyTGTR) schematic diagram and network structure

data (text descriptions of events) and structured data (time series, weather information, and additional event information). The output is the predicted result under this input.

Specifically, the text data was represented in the C module. The text feature representation learned by each SConv1D block corresponds to the early feature representation, mid-term feature representation, and late feature representation of the text, respectively. They detected specific patterns in text at different levels of abstraction. TGTR established an interaction between unstructured textual data and data such as structured tense. TGTR was applied to learn text feature representations at an early stage and obtained "reset" textual abstract feature representations. Then, the final text representation was obtained through mid- and late-stage pattern learning.

In module E, two parts of latent abstract features from module C and module D were concatenated to obtain a joint feature representation tensor. Finally, the taxi traffic demand prediction in the event area was performed by fusing joint features at the FC layer.

Experiments

Data description

The base dataset is a large-scale public dataset of 110 million taxi trips from New York. This dataset was published by the NYC Taxi and Limousine Commission (TLC) on its website, which published individual taxi records within the city of New York from January 2009 to June 2016 [37]. On the basis of this data, a range of about 500 m near Terminal 5 [38] in the center of Manhattan was selected as the research object, and the taxi trip data in this range was explored.

The textual data for the experiment came from information about cyber incidents in the area around Terminal 5. 315 event information in similar time periods were recorded, including event date, specific time, title and corresponding text description. Part of the event information is shown in Table 1.

Table 1 Event information acquired from internet

Title	Description
Walk the Moon at Terminal 5 on 4/14	WALK THE MOON at Terminal 5 on 4/14 (Sold Out) All Ages
Local Natives & with Charlotte Day Wilson	Charity: Local Natives believe in equality, safety, and dignity of all people. They have partnered with Plus 1 so that $1 from every ticket is going to support gender-based violence intervention…
Ringling Bros. Circus	Witness the Greatest Show On Earth one last time! Be a part of the last curtain call when Ringling Bros. and Barnum & Bailey presents Out Of This World, coming to Barclays Center…
Arcade Fire	Due to overwhelming demand, Grammy Award-winning band, Arcade Fire, announced additional dates for the highly-anticipated REFLEKTOR TOUR in support of its international #1 album…
2014 NBA Playoffs—Nets vs. Heat -Game 3	Its Game 3 of the second round of the NBA Playoffs, Saturday, May 10 when the Brooklyn Nets take on the Miami Heat at Barclays Center…

Table 2 Part of the weather data for 10 consecutive days

Date	Min-temp(F)	Max-temp(F)	Rain-drizzle	Precipitation(mm)	Snow-ice
2013/1/10	39.9	48	0	0	0
2013/1/11	37	44.1	1	0	0
2013/1/12	42.1	46.9	1	0.57	0
2013/1/13	43	48.9	0	0.02	0
2013/1/14	48	72	0	0	0
2013/1/15	35.1	50	1	0.09	0
2013/1/16	32	37.9	1	0.63	1
2013/1/17	36	43	0	0.09	0
2013/1/18	25	39.9	0	0	0
2013/1/19	28	51.1	0	0	0

In addition, events that occurred in different time periods had different degrees of influence on the demand for taxis the next day. For example, an unexpected event in the middle of the night may lead to a surge in demand for taxis the next day, so depending on the time period of the event, an additional event feature was added, that is, whether there was an event in that time period.

At the same time, considering the influence of weather information on taxi travel in real life, factors such as rain, snow, temperature, etc. will change people's travel mode to a certain extent, so these features contain some weather characteristics. Weather data came from measurements at New York's Central Park Weather Station. As shown in Table 2, partial weather information for 10 consecutive days from January 10, 2013 to January 19, 2013 was displayed.

Hyperparameters and loss function

The training of the entire end-to-end neural network is done in the Keras environment. The mean squared error (MSE) was chosen as the training loss function, defined by Eq. (1). The RMSprop (lr = 0.001, roh = 0.9, epsilon = 1e-08) optimizer was used. Backpropagation was performed on a mini-batch of size 64 to train the fusion prediction network. Each training was performed for 700 iterations and the optimal weights of the validation set during

the iterations were retained. In addition, some locations on the network also added dropout (Dropout ratio = 0.5). L2 regularization (L2 = 0.05) was added to the nonlinear FC layer. These works were mainly aimed at reducing overfitting.

$$loss(MSE) = \frac{1}{M}\sum_{m=1}^{M}(y_m - \widehat{y}_m)^2 \tag{1}$$

In the formula, y_m represents true value of the samples in the training set; $\widehat{y}_m$ represents the predicted value of the samples in the training set; M is the number of samples in the training set.

Performance metrics

In order to evaluate the performance of various models, the regression error statistics of the test set were performed from the following perspectives, including mean absolute error (MAE), relative absolute error (RAE), root mean square error (RMAE), relative root square error (RRSE), mean absolute percentage error (MAPE) and judgment coefficient (R2). The specific formula is shown in Table 3.

Among them, y_n denotes the true value of samples in testing dataset; $\widehat{y}_n$ denotes the predicted value of samples in testing dataset; $\overline{y}_n$ denotes the mean of the true values of samples in testing dataset; N denotes the number of samples in testing dataset.

Implementation

Time series detrending

Detrending the nonlinear non-stationary data is an important step in data analysis. There is a clear cyclical trend in traffic flow prediction, which is caused by daily habitual behavior. Helps improve forecasting performance by eliminating repetitive trends contained in the data. First, the historical average for each day of the week was calculated based on historical data, as in Eq. 2. Historical averages represented a fixed cyclical trend. Then, the data was centered by the historical mean, and then scaled by the standard deviation, as shown in Eq. 3. The goal of a predictive model was to learn to predict residuals resulting from detrending.

$$\overline{x}_i = \frac{1}{D}\sum_{d=1}^{D} x_i \tag{2}$$

Table 3 Evaluation index

$MAE = \frac{1}{N}\sum_{n=1}^{N}\left\|y_n - \widehat{y}_n\right\|$	$RAE = \frac{\sum_{n=1}^{N}\left\|\widehat{y}_n - y_n\right\|}{\sum_{n=1}^{N}\left\|y_n - \overline{y}_n\right\|}$	$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n - \widehat{y}_n\right)^2}$
$RRSE = \frac{\sqrt{\sum_{n=1}^{N}\left(y_n - \widehat{y}_n\right)^2}}{\sum_{n=1}^{N}(y_n - \overline{y}_n)^2}$	$MAPE = \frac{100}{N}\sum_{n=1}^{N}\left\|\frac{y_n - \widehat{y}_n}{y_n}\right\|$	$R2 = 1 - \frac{\sum_{n=1}^{N}\left(y_n - \widehat{y}_n\right)^2}{\sum_{n=1}^{N}(y_n - \overline{y}_n)^2}$

$$x_i' = \frac{x_i - \overline{x}_i}{\sigma} \tag{3}$$

In the formula, $i \in [0,6][0,6]$, 0 to 6 represent Monday to Sunday; D represents the number of samples of i in the historical data; σ represents the overall standard deviation.

Experimental design

In order to evaluate the effects of several key components of FFN, "ablation" experiments were designed to quantify the influence of different components on the fusion prediction results. Key components of FFN were replaced or removed while keeping other parameters constant. Secondly, in order to quantify the degree of influence of different data sources on the prediction results, "incremental" studies were carried out, that is, adding different data sources in sequence to the input of the prediction network. Finally, the following sets of models were given.

(1) TGTR was embedded in the early text features, mid-term text features and late text features of the fusion prediction network, respectively, and other network structures remained unchanged. FFN(SConv1D + earlyTGTR), FFN(SConv1D + middleTGTR) and FFN(SConv1D + lateTGTR) were designed to explore the effect of TGTR on textual information at different stages.
(2) FFN (SConv1D + none) was a sub-network that removed TGTR and kept other structures unchanged, where none represented that TGTR was not added. This was done to assess the effects of TGTR components.
(3) FFN(Conv1D + none) replaced the SConv1D layer with a regular one-dimensional convolutional layer (Conv1D).
(4) The advanced model mentioned in [19].
(5) The influence of different information sources on the prediction results was evaluated in the FFN (SConv1D + earlyTGTR) network. L represents time series information, W represents weather feature information, E represents information about the existence of an event, and T represents text information describing the event.
(6) FFN was compared with three popular time series forecasting methods, support vector regression (SVR), gaussian process (GP) regression, and autoregressive integrated moving average (ARIMA).

Performance evaluation and comparison

The processed data was divided into three parts: training set, validation set and test set. The data of 2013 and 2014 is the training set. The data in 2015 is the validation set. The remaining data from 2016 was used for testing. The corresponding sample numbers are 730, 365, and 170, respectively. The test results of different methods on the test dataset are shown in Table 4.

From the perspective of various evaluation indicators, the FFN (SConv1D + earlyTGTR) network has the best performance. Among all methods, the FFN (SConv1D + earlyTGTR) network has the lowest MAE with relatively few training parameters. It can be seen that the TGTR application has the best prediction performance in the early text representation. This is mainly due to the guidance of data such as temporality, which

Table 4 Comparison of the performance of different methods on the test data set

Method(L + W + E + T)	MAE	RAE (× 100)	RMSE	RRSE (× 100)	MAPE	R2 (× 100)	Trainable params
FFN(SConv1D + earlyTGTR)	148.7	63.0	227.7	67.8	15.5	54.1	198101
FFN(SConv1D + middleTGTR)	150.0	63.5	228.6	68.0	15.7	53.7	197641
FFN(SConv1D + lateTGTR)	152.1	64.4	231.0	68.7	16.0	52.8	197641
FFN(SConv1D + none)	151.6	64.2	229.3	68.2	15.9	53.4	196911
FFN(Conv1D + none)	154.5	65.4	230.2	68.5	16.4	53.1	205821
DL-LSTM [19]	160.8	–	233.5	–	16.7	51.7	–
DL-FC [19]	152.6	64.6	232.1	69.1	16.1	52.3	238313

Table 5 TFFN (SConv1D + earlyTGT) performance on the test set under different information sources

FFN(SConv1D + earlyTGTR)	MAE	RMSE	R2(× 100)
L	188.1	254.2	42.8
L + W	184.8	252.9	43.4
L + W + E	168.9	242.9	47.8
L + W + E + T	148.7	227.7	54.1

enables the model to filter secondary features well and capture the most salient text representations at an early stage. In addition, the effect of TGTR applied in the mid-late text representation is not as good as that in the early text representation, and the prediction results are gradually worse. By comparing the results of FFN (SConv1D + lateTGTR) and FFN (SConv1D + none), we find that TGTR has a disturbing effect on the later text representation.

As can be seen from Table 4, in the absence of TGTR components, SConv1D has great advantages over conventional Conve1D. Compared with the Conve1D network, the prediction performance of this model is improved, and the MAE is reduced from 154.4 to 151.6. Furthermore, the evaluation results of FFN (SConv1D + earlyTGTR) outperform other state-of-the-art prediction methods. Compared with the DL-LSTM model, the optimal model FFN (SConv1D + earlyTGTR) proposed in this paper reduces the MAE by 12.1, and the prediction performance is significantly improved. Although the MAE of FFN (SConv1D + earlyTGTR) is only 3.9 lower than that of the DL-FC model, the number of training parameters is reduced by 40,212. The performance of FFN(SConv1D + earlyTGTR) on the test set does not lead to insufficient model fitting effect due to the reduction of training parameters. This is the advantage of FFN (SConv1D + earlyTGTR) over the DL-FC model.

In order to evaluate the contribution of different information sources, different information sources were sequentially added to the network. The results are shown in Table 5. Each time a modal information is added, the prediction performance of the model is improved to varying degrees, especially after adding event information, the MAE of the model reaches 148.7. The experimental results show that compared with the prediction network that only considers traffic data, the fusion prediction network reduces MAE by 26.5%, RMSE decreases by 11.6%, and R2 increases by 26.4%. The textual information

of events in different modal information sources contributed the most to the improved model. These results clearly highlight the importance of fusing time-series data with textual informative data, especially for the taxi traffic demand forecasting problem in the event region considered in this paper.

Figure 4 shows the prediction curves of network test samples under different input information. As can be seen from the figure, considering the textual information describing the event in the prediction network can improve the prediction accuracy. This is consistent with real scenarios, as textual information often contains contextual interpretations of the multimodalities observed in temporal data. For example, if a stadium is preparing for a football game, taxi traffic in the area may be much higher than usual. The fusion prediction network considers the impact of special events and dynamically predicts traffic demand.

Furthermore, FFN (SConv1D + earlyTGTR) was compared with SVR and GP regression (Fig. 5), two popular time series forecasting methods. The SVR method used two kinds of kernel functions, linear kernel and radial basis function (RBF). The hyperparameters of SVR performed an exhaustive search using all possible combinations of the set {0.001, 0.01, 0.1, 1.0, 10, 100}. The randomly selected hyperparameters were brought into the network for training and testing, and the optimal model obtained by the search was saved with the optimal accuracy of the validation set as the indicator, and then the corresponding hyperparameters were returned. GP regression used the RBF kernel function. The corresponding hyperparameters of SVR and GP are shown in Table 6.

Finally, the method in this paper was compared with the ARIMA model. In ARIMA (p, d, q), AR is "autoregression", p is the number of autoregressive items; MA is "moving average", q is the number of moving average items; d is the number of differences made to make it a stationary series (order). This paper sets the option set of ARIMA hyperparameters as follows: $p \in \{0, 1, 2, 3, 4\}$, $q \in \{0, 1, 2, 3, 4\}$, $d \in \{0, 1, 2\}$. The minimum bayesian information criterion (BIC) is used as the evaluation index to select the optimal hyperparameters of the ARIMA model, and the results are shown in Fig. 6.

As can be seen from the figure, under the training set, the BIC of the ARIMA ($p = 3$, $d = 1$, $q = 1$) model is the lowest. Test results of ARIMA ($p = 3$, $d = 1$, $q = 1$) model:

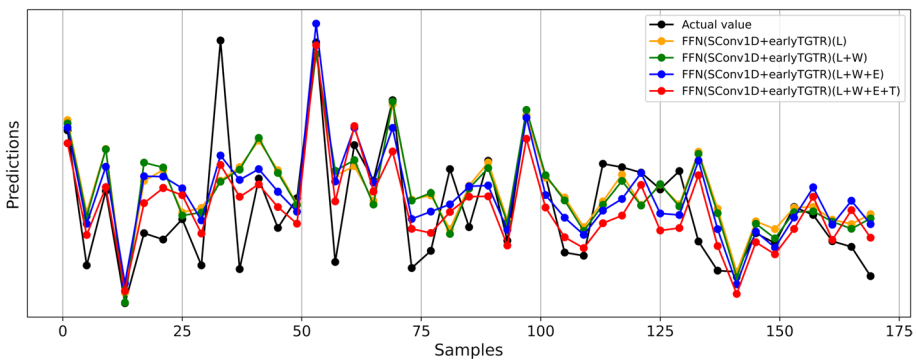

Fig. 4 Prediction results of the FFN (SCon1D + earlyTGT) model on the test set with different information sources

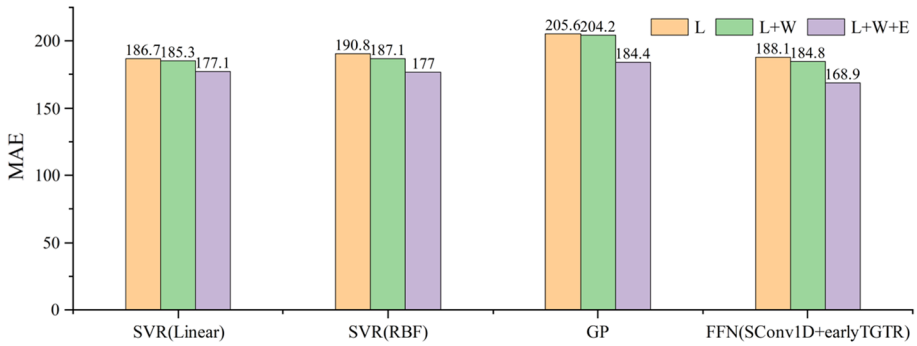

Fig. 5 Incremental experiments with different methods

Table 6 Hyperparameter settings

Method	Kernel	Hyperparameters
SVR	Linear	C = 100
	RBF	C = 10 gamma = 0.01
GP	RBF	length_scale = 1
		length_scale_bounds = (0.01, 1000)

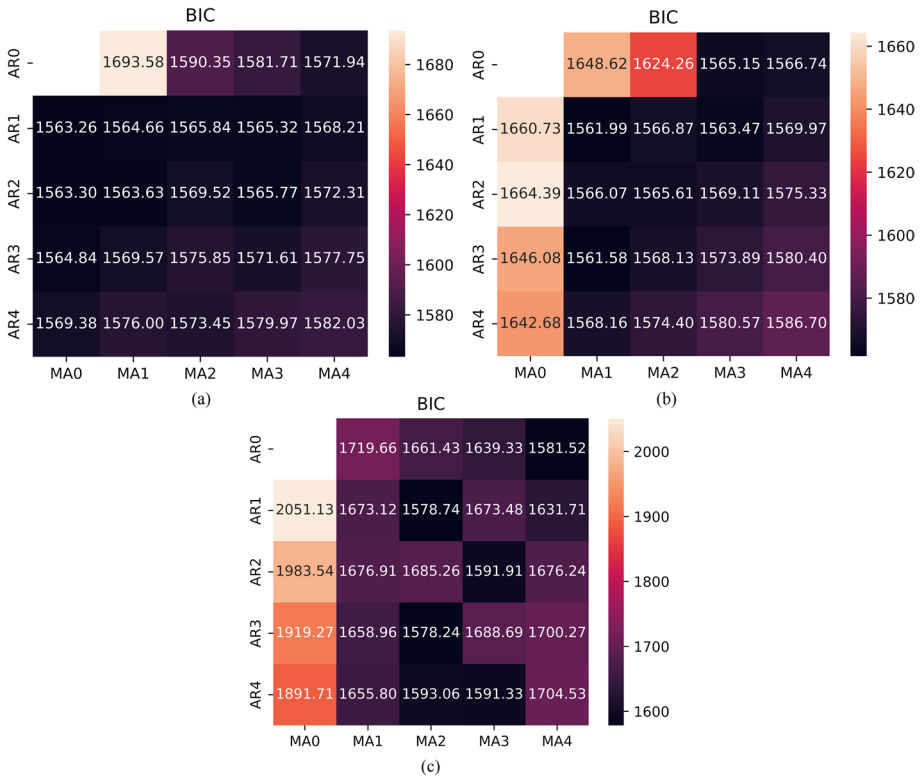

Fig. 6 BIC of ARIMA model with different hyperparameters **a** d = 0; **b** d = 1; **c** d = 3

MAE is 167.2, RMSE is 241.0, and R2 is 0.486. Even though the ARIMA model is carefully tuned, its MAE is still 18.5 higher than the best results of our FFN, and its RMSE is 13.3 higher. It can be seen that the method proposed in this paper has strong competitiveness. Meanwhile, experimental results show that considering the impact of event information is very important for predicting taxi traffic near special event areas.

Conclusion

This paper focused on the impact of event information on traffic flow demand, and proposed FFN combining special event information and traffic flow information. The lightweight SConv1D reduced the learning parameters of the model by performing depth convolution and point-by-point convolution on the three-dimensional floating-point tensors of the input text respectively; temporal and other data-driven networks were used to obtain the class probability score of word features, so as to "reset" the representation of text abstract features and apply it to each stages of text coding; a simple and easy to implement splicing and fusion strategy was used to fuse multimodal features. Extensive experimental results show that the information source that has the greatest impact on the prediction performance of different methods is the information about events; by using the textual information of regional events, the proposed deep network framework significantly outperforms other popular time series forecasting methods that do not consider textual information modeling.

In future work, we aim to explore how to use advanced deep learning techniques to allow models to better understand the various types of information describing events; develop a city-wide spatiotemporal model to account for all event information pairs occurring across the city The impact of traffic forecasts.

Acknowledgements
Not applicable.

Author contributions
Yu and Sun designed the study, conducted the experiments and analyzed the results. Liu and Zhang supervised the completion of the work and contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China under Grants 62173137, 52172403. Hunan Provincial Natural Science Foundation of China under Grants 2021JJ50001, 2021JJ30217. Project of Hunan Provincial Department of Education grant number 19A137.

Availability of data and materials
The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that there are no conflicts of interest regarding the publication of this paper.

Received: 18 December 2021 Accepted: 7 June 2022
Published online: 21 June 2022

References

1. Hao S, Zhang M, Hou A. Short-term traffic flow forecast based on DE-RBF fussion model. J Phys: Conf Ser. 2021;1910(1): 012035.
2. Fang W, Zhuo W, Yan J, et al. Attention meets long short-term memory: a deep learning network for traffic flow forecasting. Physica A. 2022;587: 126485.
3. Rajalakshmi V, Ganesh Vaidyanathan S. Hybrid CNN-LSTM for traffic flow forecasting. In: Proceedings of 2nd international conference on artificial intelligence: advances and applications. 2022. p. 407–414.
4. Valentinetti D, Muñoz FF. Internet of things: emerging impacts on digital reporting. J Bus Res. 2021;131:549–62.
5. Zhang C, Chen H, He J, et al. Reconstruction method for missing measurement data based on wasserstein generative adversarial network. J Adv Comput Intell Intell Inf. 2021;25(2):195–203.
6. Salazar-Carrillo J, Torres-Ruiz M, Davis CA, et al. Traffic congestion analysis based on a web-GIS and data mining of traffic events from Twitter. Sensors. 2021;21(9):2964.
7. Almuqren L, Alrayes FS, Cristea AI. An empirical study on customer churn behaviours prediction using arabic Twitter mining approach. Future Internet. 2021;13(7):175.
8. Zhang T, Shen S, Cheng C, et al. A topic model based framework for identifying the distribution of demand for relief supplies using social media data. Int J Geogr Inf Sci. 2021;35(11):2216–37.
9. Huang Y, Gan H, Jing P, et al. Analysis of park and ride mode choice behavior under multimodal travel information service. Transp Lett. 2021: 1–11.
10. Tian H, Tao Y, Pouyanfar S, et al. Multimodal deep representation learning for video classification. World Wide Web. 2019;22(3):1325–41.
11. Jiang Y-G, Wu Z, Wang J, et al. Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Trans Pattern Anal Mach Intell. 2018;40(2):352–64.
12. Chen Q, Wang W, Huang K, et al. Multi-modal generative adversarial networks for traffic event detection in smart cities. Expert Syst Appl. 2021;177: 114939.
13. Chen Q, Wang W. Multi-modal neural network for traffic event detection. In: 2019 IEEE 2nd international conference on electronics and communication engineering. 2019. p. 26–30.
14. Wen H, You S, Fu Y. Cross-modal context-gated convolution for multi-modal sentiment analysis. Pattern Recogn Lett. 2021;146:252–9.
15. Zhang D, Li S, Zhu Q, et al. Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning. IEEE Access. 2020;8:22945–54.
16. Linqin C, Zhongxu L, Sitong Z et al. Visual question answering combining multi-modal feature fusion and multi-attention mechanism. In: 2021 IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering. 2021. p. 1035–1039.
17. Xiao W, Kuang L, An Y. Traffic flow prediction through the fusion of spatial-temporal data and points of interest. In: International conference on database and expert systems applications. 2021. p. 314–327.
18. Yu K, Qin X, Jia Z, et al. Cross-attention fusion based spatial-temporal multi-graph convolutional network for traffic flow prediction. Sensors. 2021;21(24):8468.
19. Rodrigues F, Markou I, Pereira F-C. Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach. Information Fusion. 2019;49:120–9.
20. Li W, Chen S, Wang X, et al. A hybrid approach for short-term traffic flow forecasting based on similarity identification. Mod Phys Lett B. 2021;35(13):2150212.
21. Lu S, Zhang Q, Chen G, et al. A combined method for short-term traffic flow prediction based on recurrent neural network. Alex Eng J. 2021;60(1):87–94.
22. Pereira FC, Rodrigues F, Polisciuc E, et al. Why so many people? Explaining nonhabitual transport overcrowding with internet data. IEEE Trans Intell Transp Syst. 2015;16(3):1370–9.
23. Chen X, Lu J, Zhao J, et al. Traffic flow prediction at varied time scales via ensemble empirical mode decomposition and artificial neural network. Sustainability. 2020;12(9):3678.
24. Kuppam A, Copperman R, Rossi T, et al. Innovative methods for collecting data and for modeling travel related to special events. Transp Res Rec J Transp Res Board. 2011;2246(1):24–31.
25. Chang M-S, Lu P-R. A multinomial logit model of mode and arrival time choices for planned special events. J East Asia Soc Transp Stud. 2013;10:710–27.
26. Noursalehi P, Koutsopoulos H-N, Zhao J. Real time transit demand prediction capturing station interactions and impact of special events. Transp Res Part C Emerg Technol. 2018;97:277–300.
27. Yao W, Qian S. From Twitter to traffic predictor: next-day morning traffic prediction using social media data. Transp Res Part C Emerg Technol. 2021;124: 102938.
28. Gao J, Li P, Chen Z, et al. A survey on deep learning for multimodal data fusion. Neural Comput. 2020;32(5):829–64.
29. Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing. 2017. p. 1103–1114.
30. Fukui A, Park D-H, Yang D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 conference on empirical methods in natural language processing. 2016. p. 457–468.
31. Liu Z, Shen Y, Lakshminarasimhan V-B, et al. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol 1. 2018. p. 2247–2256.
32. Wu A, and Han Y. Multi-modal circulant fusion for video-to-language and backward. In: Twenty-seventh international joint conference on artificial intelligence, vol 3(4). 2018. p. 8
33. Guo W, Wang J, Wanga S. Deep multimodal representation learning: a survey. IEEE Access. 2019;7(99):63373–94.
34. Church K-W. Word2Vec. Nat Lang Eng. 2017;23(1):155–62.

35. Pennington J, Socher R, Manning C-D. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. 2014. p. 1532–1543.
36. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1251–1258.
37. New York City Taxi & Limousine Commission. Taxi and limousine commission trip record data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed 13 Dec 2021.
38. Terminal 5. https://www.nyc.com/bars_clubs_music/terminal_5.993530/. Accessed 13 Dec 2021.

Publisher's Note