Deep Dynamic Fusion Network for Traffic Accident Forecasting

Chao Huang[1], Chuxu Zhang[2], Peng Dai[1], Liefeng Bo[1]
JD Digits[1]
University of Notre Dame, USA[2]
chaohuang75@gmail.com,czhang11@nd.edu,{peng.dai, liefeng.bo}@jd.com

ABSTRACT

Traffic accident forecasting is a vital part of intelligent transportation systems in urban sensing. However, predicting traffic accidents is not trivial because of two key challenges: i) the complexities of external factors which are presented with heterogeneous data structures; ii) the complex sequential transition regularities exhibited with time-dependent and high-order inter-correlations. To address these challenges, we develop a deep Dynamic Fusion Network framework (DFN), to explore the central theme of improving the ability of deep neural network on modeling heterogeneous external factors in a fully dynamic manner for traffic accident forecasting. Specifically, DFN first develops an integrative architecture, *i.e.*, with the cooperation of a context-aware embedding module and a hierarchical fusion network, to effectively transferring knowledge from different external units for spatial-temporal pattern learning across space and time. After that, we further develop a temporal aggregation neural network layer to automatically capture relevance scores from the temporal dimension. Through extensive experiments on real-world data collected from New York City, we validate the effectiveness of our framework against various competitive methods. Besides, we also provide a qualitative analysis on prediction results to show the model interpretability.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**.

KEYWORDS

Traffic accident forecasting; Spatial-temporal prediction; Deep learning; Intelligent transportation

ACM Reference Format:
Chao Huang[1], Chuxu Zhang[2], Peng Dai[1], Liefeng Bo[1]. 2019. Deep Dynamic Fusion Network for Traffic Accident Forecasting. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3357384.3357829

1 INTRODUCTION

Traffic is the pulse of a city that impacts our daily life. One of the most fundamental challenges in building intelligent transportation systems is accurate traffic accident forecasting, which provides useful information for decision-making and public safety [3, 30]. For example, by predicting future citywide traffic accidents, (i) governments can design better transportation planning and scheduling strategies to alleviate traffic accidents and congestion beforehand [35, 39]; and (ii) it can help people evaluate the risk of potential motor vehicle collision and handle crowd aggregations in an efficient way [14, 24]. Hence, it has become a pressing need to understand the occurrence patterns of traffic accidents at geographical regions in a city from their historical records and foresee future collision occurrences.

A number of studies have investigated traffic accident forecasting [23, 27]. However, most of these studies identify the traffic accidents by analyzing the historic traces or movement patterns of the studied objects using statistical and traditional linear machine learning approaches (*e.g.*, SVM [27] and ARIMA [2]). In recent years, neural network models, such as recurrent neural network (RNN), has been utilized for modeling spatial-temporal sequences with non-linearites, which achieves noticeable improvements over traditional methods [19, 37**?**]. Despite their effectiveness, we argue that these methods are not sufficient to yield satisfactory forecasting performance on traffic accidents. In fact, traffic accidents can only happen at a small number of regions across the entire city. In the scenario of modeling traffic accidents, imbalanced data in the generated traffic accident sequence is harmful to deep learning models–largely relying on data sufficiency [47]. The finer the granularity for geographical region discernment, making the already imbalanced data even worse [48]. Therefore, to build effective predictive models with such imbalanced traffic accidents, it is crucial to account for the influences from different external factors, so as to transfer knowledge from different data sources to alleviate data scarcity in the forecasting task.

However, to solve the traffic accident forecasting problem, we face two key challenges. First, the occurrences of traffic accidents are affected by various external factors, *e.g.*, the distribution of region function in urban space, and citywide abnormal events [44]. These factors can be either stationary and dynamic. For example, Point-of-Interest (POI) data features the function of each geographical region and is helpful to model latent correlation between different regions [13, 38]. In addition, abnormal events (*e.g.*, blocked driveway, traffic jam and road construction) happening in real-time could reflect city dynamics and provide new insights to advance our understanding of traffic accident occurrences in a dynamic manner. Hence, how to effectively preserve the heterogeneous structures (stationary and dynamic) of external factors in our predictive framework remains a significant challenge.

In addition, the inter-dependencies between traffic accidents and external factors could be dynamic (*i.e.*, changing from time to time) in practice [18]. For instance, a traffic accident from the

CIKM'19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357829

same business center may be correlated with the road construction during the weekday, or the large-scale entertainment events held on weekend. Thus, a flexible solution that is capable of adapting the knowledge learned from external sources dynamically is required. Furthermore, it is not necessary that future traffic accident occurrence will be more relevant to a recent one than one that is further away. It is promising to endow the traffic accident forecasting framework with the capability to capture both short-term and long-term dependencies.

To address the above challenges, we propose a novel framework, namely deep Dynamic Fusion Network framework (DFN), for traffic accident forecasting. In the architecture of DFN, we first introduce a context-aware embedding module to retain the fine-grained semantics of cross-region relations and facilitate the emebdding learning on geographical areas. Then, we develop a hierarchical fusion network to model the inter-dependencies between traffic accidents and other external factors with a multi-level hierarchy of recurrent architecture. The proposed hierarchical fusion network allows external knowledge to guide the cross-modal pattern representation learning process. Additionally, another key component in DFN is a temporal aggregation mechanism, which aims to generate conclusive latent representations in a principled way. This aggregation module is jointly trained to capture the relevance between the current time step and previous ones. The contributions of this work are as recapitulated as follows:

- We propose a new DFN framework to predict traffic accidents with the careful exploration of heterogeneous external data sources. To the best of our knowledge, this is the first framework that is designed to tackle the problem of fine-gained traffic accident forecasting by fusing both stationary and dynamic influence of external factors.
- We develop a hierarchical fusion network to aggregate the external factors in a fully dynamic manner. With the cooperation of the hierarchical structured recurrent framework and context-aware embedding module, our DFN could promote the collaboration of different views (spatial, temporal and semantic) for accurate traffic accident forecasts. In addition, a temporal aggregation mechanism is proposed to automatically capture the importance of each time-specific representation in the predictive model.
- We have conducted extensive experiments on the real-life datasets from New York City. Our experimental results show that, compared with state-of-the-art methods, DFN can better transfer knowledge from relevant heterogeneous external sources and achieves better performance for traffic accident predictions.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. We formulate the studied traffic accident forecasting problem and discuss relevant external sources in Section 3. Following the motivation to address the challenges, we present the details of the architecture of DFN in Section 4. In Section 5, we perform extensive experiments to evaluate the model performance. Finally, we conclude our paper in Section 6.

2 RELATED WORK

In this section, we review the research work which are relevant to our work from the following perspectives: traffic accident analysis and forecasting, spatial-temporal predictive models. We also discuss the difference between this work and the related work.

2.1 Traffic Accident Analysis and Forecasting

Traffic accident detection and forecasting have become fundamental challenges in urban sensing. Prior work have made significant advances on the analysis of traffic accidents from different dimensions [7, 22, 34]. For example, a new spatial clustering framework was developed for risk mapping with the consideration of traffic accident severity levels [34]. Deb *et al.*[7] proposed a new imputation method to infer missing values of historical traffic accident. Additionally, some other work are proposed to predict traffic accidents by considering various correlations in urban space [27, 45].

In general, work on traffic accident forecasting can be grouped into two categories: conventional pattern-based methods and deep neural network models. For example, Bharti *et al.*[27] developed a support vector machine based approach to infer future traffic accidents. Chong *et al.* [23] used decision tree and neural network to model relations between accidents and driver's behavior. However, pattern-based methods mostly rely on stationary assumption of the traffic accident data. Hence, many follow-up works are introduced to capture dynamic traffic changes using deep learning techniques. For example, Ren *et al.* [26] extened the RNN architecture to uncover the future traffic accident risk. Yuan *et al.* [45] solved the traffic accident prediction problem using the convolutional long short-term memory neural network to model the correlations between regions over time. Different from those work, our DFN framework targets at fine-grained traffic accident forecasting by not only modeling both spatial-temporal dependencies, but also automatically aggregating heterogeneous external factors in a dynamic manner.

2.2 Spatial-Temporal Predictive Models

We review existing research work in spatial-temporal forecasting [15, 40, 42, 49]. This line of work is also relevant with our studied traffic accident forecasting problem, which is also a sequential data forecasting problem with spatial and temporal information. A number of earlier studies have investigated spatial-temporal prediction using conventional approaches. For example, Matias *et al.* [21] applied autoregressive integrated moving average (ARIMA) in traffic prediction. Vector Auto-Regressive (VAR) was used to model linear inter-dependencies of space and time-ordered data [36]. However, these traditional methods are mostly limited to linear models which may not perform well for dynamic spatial-temporal patterns.

Recent advances in deep learning has inspired researchers to model non-linear spatial and temporal correlations with deep neural network techniques. Recurrent neural networks (RNNs) have been successfully applied in many spatial-temporal forecasting related tasks, such as human mobility modeling [28], taxi demand prediction [40], forecasting traffic volume [42] and location recommendation [49]. In particular, Yao *et al.* [40] developed an integrative framework of convolutional and recurrent neural networks to infer future taxi demand. Song *et al.* [28] developed a deep neural network transportation system for modeling human mobility based on the collected heterogeneous data source. Yu *et al.* [42] proposed a stacked LSTM network to forecast peak-hour traffic with the consideration of both normal and extreme traffic conditions. Zhao *et al.* [49] studied the POI recommendation task with a context-aware embedding learning method. Additionally, different attention mechanisms are introduced to improve the performance of RNNs in forecasting spatial-temporal data. For example, attention mechanism is utilized to preserve the sequential information among historical records in modeling human mobility traces [9]. In this work, we design a temporal-wise recalibration mechanism which is tailored to cooperate with hierarchically structured recurrent framework, to adaptively guide the networks to focus on particular parts of aggregated representations.

3 PRELIMINARY AND PROBLEM STATEMENT

In this section, we briefly describe key notations. Then, we formally present the traffic accident forecasting problem in this paper.

3.1 Notations

Given a window of T time steps (*e.g.*, day), we suppose there are M (indexed by m) geographical regions, each of which (*i.e.*, region R_m) corresponds to a target temporally ordered sequences of traffic accident occurrences, *i.e.*, $Y_m = \{y_m^1, ..., y_m^t, ..., y_m^T\}$. In Y_m, the element y_m^t is set to 1 if there exist traffic accidents happened at the region R_m in the t-th time step, and $y_m^t = 0$ otherwise. In addition, we further define the citywide event data source as external factors and represent it as $\Lambda_m = \{\lambda_m^1, ..., \lambda_m^t, ..., \lambda_m^T\}$ denotes the abnormal event occurrences of region R_i over T time steps. In addition to citywide abnormal events, Point-of-Interests information is also useful for capturing the correlations between different geographical regions while making forecasting on traffic accidents, which advances the modeling of semantic signals from spatial dimension [43]. For example, geographical regions which have similar function (*e.g.*, business center or tourist attraction) in urban areas may share similar occurrence patterns of traffic accidents. In particular, each geographical region is associated with a vector p_m which contains the statistical information of G venue categories. We define a POI Matrix $\mathcal{P} \in \mathbb{R}^{M \times G}$ to represent the POI information of all regions $R_m, m \in [1, ..., M]$ over G venue categories.

3.2 Problem Statement

The objective of the traffic accident forecasting is that, given observations of historical traffic accidents ($Y_m, m \in [1, ..., M]$) and external data sources (abnormal events–$\Lambda_m, m \in [1, ..., M]$ and point-of-interests– $\mathcal{P}$), how to predict traffic accident occurrences in the future time step (*i.e.*, Y_m^{T+1}).

$$Y_m^{T+1} = F(Y_m, \Lambda_m, \mathcal{P}); \quad (m \in [1, ..., M]) \tag{1}$$

where $F(\cdot)$ is the mapping function we aim to learn.

4 METHODOLOGY

In this section, we will explain the details of our developed DFN framework for traffic accident forecasting.

4.1 Region Embedding Module

Intuitively, locations sharing similar functionality may have similar traffic patterns which might be relevant to regions' accident

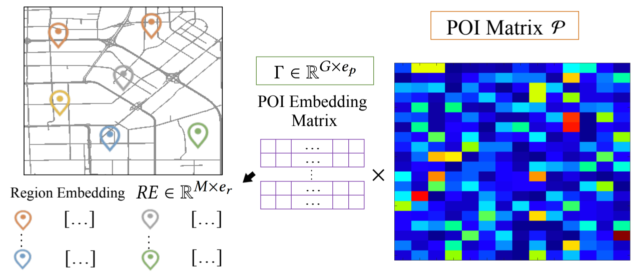


Figure 1: The Region Embedding Module.

occurrences. While there exist research work modeling region correlations using convolutional neural network, they only handle the local structure similarity among adjacent geographical regions [40]. Such consideration cannot deal with the region correlations in a global perspective. To encode the semantic signals into the region-wise correlation learning process, we incorporate the POI information into the hidden representation generation of the geographical region (as shown in Figure 1). To be more specific, we first define a point-of-interest embedding matrix $\Gamma \in \mathbb{R}^{G \times e_p}$ which map each point-of-interest category into a hidden representation vector. Then, we generate a new matrix Φ by performing the operation of $\Phi = \mathcal{P} \cdot \Gamma$. The signals of point-of-interest will be captured in the learning process via computing the difference between the region embedding matrix $RE \in \mathbb{R}^{M \times e_r}$ and Φ, *i.e.*, $|RE - \Phi|$. The embedding vector of each region R_m is denoted by RE_m.

4.2 Pattern Fusion Module

In the pattern fusion module of our DFN framework, we aim to encode the sequential patterns of traffic accident occurrences as well as the correlations between traffic accidents and external factors. For each region R_m, we generate a traffic occurrence vector O_m^T (with length L) to reflect the occurrences of region R_m's previous L traffic accidents and each element $O_{m,l}^T$ is the concatenation of region embedding RE_m and the embedding vector TE_l corresponding to the occurrence time information of l-th traffic accident at region R_m. Following the similar way, we could generate O_m^E to indicate the occurrences of region R_m's previous L reported abnormal events. The time embedding is constructed with a fix (non-trainable) date embedding using the timing signal method [32].

Given the generated two temporally ordered vectors O_m^T and O_m^E of region R_m, we propose to leverage recurrent neural networks (*e.g.*, Gated Recurrent Units (GRU) [6] and Long Short-Term Memory (LSTM) [11]) to encode the sequential patterns. Our DFN framework is flexible to the recurrent unit architecture and we consider GRU as a concrete example in this work. GRU has been developed as an improved version of recurrent neural network with gating mechanism by addressing the vanishing gradient problem [16, 33]. By utilizing the recurrent neural architecture, we could learn latent representations which preserve the ordered sequential patterns of traffic accident occurrences and relevant abnormal events. There are two gate in the GRU architecture, *i.e.*, update gate determines how much of the previous memory to pass through, and reset gate aims to learn how to combine the new input with the previous memory. Specifically, the update gate in the GRU cell is calculated as:

$$z_t = \sigma(W^z \cdot x_t + U^z \cdot h_{t-1} + b^z) \tag{2}$$

where x_t is the input into the network unit and h_{t-1} represents the hidden state holding the information from previous $t-1$ step. W^z and U^z are learnable parameters. We add the weighted information together and feed it into a sigmoid function σ to generate a value in the range of $(0, 1)$. The reset gate is computed as below:

$$r_t = \sigma(W^r \cdot x_t + U^r \cdot h_{t-1} + b^r) \tag{3}$$

After that, the current memory state is updated as follows:

$$h'_t = tanh(W^h \cdot x_t + r_t \odot U^h \cdot h_{t-1}) \tag{4}$$

$\odot$ denotes the element-wise multiplication between $U^h \cdot h_{t-1}$ and r_t. The current unit is updated based on the update gate z_t and the memory content h'_t as follows:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{5}$$

the z_t and h_{t-1} is combined with an element-wise multiplication. Similar operation is applied between $(1 - z_t)$ and h'_t. Given the generated occurrence vectors of both traffic accidents O^T_m and abnormal events O^E_m, we leverage the recurrent architecture to encode their sequential patters and learn the underlying hidden state representations for each time step. Here, we define a_t and b_t to represent the hidden state corresponding to traffic accident occurrence O^T_m and external anomaly factor O^E_m, respectively.

4.3 Temporal Aggregation Module

For a specific region, complex dependencies exist among its sequential traffic occurrences. In order to avoid the situation that only the last hidden vector is utilized to represent the temporal patterns in recurrent neural networks [25], in our DFN framework, we design the temporal aggregation mechanism to refer back to the input sequence and select the most informative signals from the time-ordered representations encoded from the pattern fusion module, which addresses the problem of recurrent neural networks in dealing with such long-range dependencies [1], namely, forgetting the earlier information of the input sequence. To achieve our goal, we propose to leverage the attention mechanism in our temporal aggregation module to look over all the information generated from our developed sequence encoder.

Attention mechanism has been widely used in many neural network-based tasks (*e.g.*, machine translation [20] and image analysis [41]) to selectively focus on parts of the input sources during the process of knowledge representation learning. Motivated by this, we build an attentive module to recalibrate the learned latent representations for traffic accident occurrence patterns. Specifically, we aim to generate a new hidden state ϱ_t which is a weighted sum of the encoder hidden states $[a_{t-1}; b_{t-1}]$, where $[\cdot;\cdot]$ represents the concentration operation. We formally define it as blow:

$$\varrho_t = (\tau_1 \cdot [a_1; b_1], \tau_2 \cdot [a_2; b_2], ..., \tau_T \cdot [a_T; b_T]). \tag{6}$$

where τ_1 represents the amount of attention we should pay to the t-th input hidden state. In our framework, τ_t is estimated over the attentive scores c_t of each individual input state c_t. We perform multi-head attention to linearly projects hidden states into d subspaces and produce the output representations via utilizing d attention functions, *i.e.*, $\Psi(\tau_t) = [head_1, head_2, ..., head_d] \cdot W_d$, where W_d is the learnable parameter. Here, we consider the scaled dot-product attention [32] as the attentive function. The learned attentive weights scores how relevant of each previous hidden state for the predicted time step. By utilizing our aggregation module over temporal dimension, it is possible for our DFN architecture to take advantage of all the intermediate hidden states and consider global temporal contextual signals, instead of solely making traffic occurrences predictions based on the last hidden state.

4.4 Model Optimization

In the forecasting phase of our proposed DFN framework, we utilize the Multilayer Perceptron (MLP) [10, 31] as the classifier to generate the occurrence probability of traffic accident of each individual region r_i. The MLP module could capture the non-linear correlations between the learned hidden representations and accident occurrence probabilities with a class of feedforward neural network, *i.e.*, $\psi_n = \phi(W_n \cdot \eta_n + b_n)$, where n denotes the index of N hidden layers in MLP module. The learned parameters are represented by W_n and b_n. To handle the non-linearities, we use *ReLU* ($\phi(\cdot)$) as the activation function. Furthermore, we further apply dropout technique [29] during the training process. To learn model parameters, we optimize the following loss:

$$\begin{aligned} \mathscr{L}_\alpha &= -\sum_{(m,t)\in S} y^t_m \ln \hat{y}^t_m + (1 - y^t_m)\ln(1 - \hat{y}^t_m) \\ \mathscr{L}_\beta &= \frac{1}{|R|} |RE - \Phi| \\ \mathscr{L} &= \mathscr{L}_\alpha + \varphi \cdot \mathscr{L}_\beta \end{aligned} \tag{7}$$

In our loss function, the estimated probability of traffic accident occurrence at region R_m and the t-th time step is indicated by $\hat{y}^t_m$.

Algorithm 1: The Model Optimization of DFN Framework.

Input: Traffic accident occurrence O^T_m, Y_m; POI matrix $\mathcal{P}$; External anomaly factor O^E_m, Λ_m; Sequence length T; and Batch size b_s.
Paras : Region embedding matrices $RE \in \mathbb{R}^{M \times e_r}$ and Other Learnable Hyperparameters θ (*e.g.*, weight matrices and bias terms).

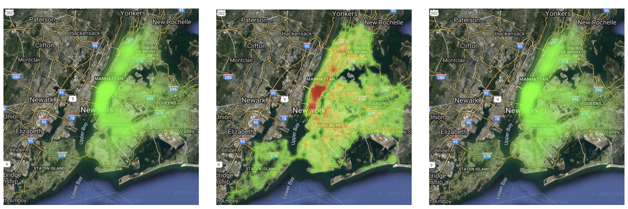
```
Initialize all parameters;
// Sample a minibatch of size b_s.
foreach T_batch = sample(X, b_s) do
    foreach ⟨m, t⟩ ∈ T_batch do
        foreach ⟨m ∈ [1, ..., M] do
            R_m = RE[(m) :];
            h^0_{m,t} = h^0 for t' ← (t − s) to (t − 1) do
                h_t = z_t ⊙ h_{t−1} + (1 − z_t) ⊙ h'_t
            end
        end
        ϱ_t = (τ_1 · [a_1; b_1], τ_2 · [a_2; b_2], ..., τ_T · [a_T; b_2T])
        ŷ_t = MLP(τ_t);
        Update loss ℒ;
    end
    Update all parameters w.r.t ℒ;
end
```

5 EVALUATION

In this section, we perform extensive evaluations and analytical experiments to verify the effectiveness of our *DFN* on the real-world datasets collected from New York City (NYC) spanning two years (2017,2018). We compare our *DFN* with many competitive baselines

Table 1: Details of the experimented datasets.

Data Source	**Citywide Traffic Accident Data**		
Time Span	Jan. 2017 - Dec. 2017	Jan. 2018 - Dec. 2018	
# of Traffic Accidents	191,554	162,995	
Data Source	**Citywide Urban Complaint Reports**		
Time Span	Jan. 2017 - Dec. 2017	Jan. 2018 - Dec. 2018	
# of Noise	413,785	442,938	
# of Blocked Driveway	127,989	134,922	
# of Illegal Parking	133,575	163,533	
# of Road Condition	80,329	105,580	
# of Unsanitary Condition	75,351	82,623	
# of General Construction	26,501	32,393	
Data Source	**Point-of-Interests (POI)**		
Category	#	Category	#
Arts & Entertainment	720	Automotive & Vehicles	1505
Business to Business	3717	Computers & Technology	637
Education	1062	Food & Dining	3385
Government	3116	Health & Beauty	4336
Home & Family	3616	Legal & Finance	1782
Real Estate	4675	Shopping	1874
Sports & Recreation	384	Others	1378



(a) Accidents on Jul. 2017 (b) Accidents on Nov. 2017 (c) Accidents on Mar. 2018

Figure 2: Geographical distributions of traffic accidents.

and show the advantages of our method with different settings. In particular, our experiments are designed to answer the following research questions:

- **RQ1**: Compared with various state-of-the-art traffic accident forecasting and spatial-temporal predictive methods, how does *DFN* perform?
- **RQ2**: How effective is our designed framework over different time periods?
- **RQ3**: How does the hierarchically structured fusion network and semantic-aware region embedding module contribute to the forecasting performance of *DFN*?
- **RQ4**: What is the impact of temporal aggregation mechanism and recurrent unit selection in the traffic accident forecasting task?
- **RQ5**: What is the influence of hyperparameters in the forecasting performance of *DFN*?
- **RQ6**: How is the interpretation of our *DFN* framework in capturing dynamic importance weights across time when forecasting future traffic accidents?

In the following parts, we will first present the experimental settings and then answer the above research questions in turn.

5.1 Experimental Settings

In this subsection, we first describe the data we use in our evaluation. Then, we introduce the compared baselines and evaluation metrics to qualify the model performance.

5.1.1 **Data Description.** Table 1 summarizes the statistical information of the real-world traffic accident dataset and two external urban data sources (*i.e.*, points of interest and urban complaint reports) collected from New York City (NYC). We further show the geographical distributions of traffic accidents in Figure 2 and present the details of each dataset as follows:

Traffic Accident Data. This dataset documents the occurrences of motor vehicle collisions from the New York Police Department. Each accident is associated with the timestamp and coordinate (latitude and longitude) information of the corresponding traffic accident. The collected traffic accident data spans two consecutive years, *i.e.*, Jan 2017 to Dec 2018.

Points of Interest Data. We collect 24,031 POIs of 14 categories (*e.g.*, Food & Dining, Shopping, Education and Real Estate & Construction) from NYC. In this dataset, each POI record contains a venue name, category and the correspondnig geo-coordinates.

Citywide Event Reports. This data is collected from the NYC's governmental non-emergency service platform which allows people in the city to report the anomalous events, by making a phone call. Each complaint record in this data includes the time, coordinates and event categories (*e.g.*, noise, road construction, blocked driveway). The time period of the collected urban complaint data is consistent with the time span of traffic accident data.

To fully evaluate the effectiveness of *DFN* framework, we adopt three widely-used metrics: *F1-score* (harmonic mean to balance precision and recall), *AUC* [8] and *Average Precision (AP)* [46]. A higher F1-score, AUC and AP value reflects a better forecasting accuracy. To ensure the fairness of comparison, our experiments are conducted by forecasting occurrences of traffic accidents in consecutive days over the test time period. The reported performance in the evaluation is averaged across all days.

5.1.2 **Compared methods.** We compare our *DFN* with the following state-of-the-art methods from various research lines, *i.e.*, conventional time series forecasting methods, feature-based learning approaches, spatial-temporal forecasting methods, attentive recurrent models and collaborative filtering-based technique.

Conventional time series forecasting methods:

- **Support Vector Regression (SVR)** [4]: it is a supervised learning model for time series regression analysis based on kernel functions, which are characterized by the margin and the number of support vectors.
- **Auto-Regressive Integrated Moving Average (ARIMA)** [2]: it is a conventional time series prediction model for understanding and predicting future values in a time series, which is in conjunction with stationary and linear transformations.

Feature-based learning approaches:

- **Logistic Regression (LR)** [12]: This learning method is a statistical model that forecasts each region's accident occurrence based on historical accident occurrences.

- **Wide and Deep Learning (Wide&Deep)** [5]: a wide & deep learning model which combines the strengths of wide linear models and deep neural networks for predictive analytics.

Spatial-temporal forecasting methods:

- **Stacked Long Short-Term Memory (S-LSTM)** [?]: it is a recurrent deep network model to consider long-term dependencies in spatial-temporal data, and aims to make predictions via jointly modeling normal and dynamic temporal patterns.
- **Spatial-Temporal Recurrent Neural Networks (ST-RNN)** [19]: it leverages the recurrent neural network for making forecasting on spatial-temporal data.

Attentive recurrent models for spatial-temporal forecasting:

- **Attentive Sequential Modeling (ASM)** [9]: it integrates RNNs with attention mechanisms to interpret the relations between the current and past values in predicting spatial-temporal data.

Collaborative filtering-based predictive technique:

- **Collaborative Filtering Predictive Technique (CFPT)** [10]: We predict traffic accident occurrences by extending this Matrix Factorization scheme with neural network architecture to handle dynamic region-time interactions.

5.1.3 **Parameter Settings.** In our experiments, we implement *DFN* with TensorFlow and train the model using Adam [17] optimizer with the learning rate of $1e^{-3}$. The hidden state dimension r is set as 32 and the input sequence length T of pattern fusion module is set as 10. In addition, we set attention size $S = 32$ and embedding dimension $e = 32$ as well. The batch size in our experiment is set to 64. All the compared approaches are trained from scratch without any pre-training on a single NVIDIA GeForce GTX 1080. We also investigate the effects of hyperparameters of *DFN* and report the evaluation results in the later subsection.

5.2 Performance Comparison (RQ1 and RQ2)

In this subsection, we first report the overall performance of our proposed *DFN* and other baselines, and then show the forecasting accuracy of all approaches over different time periods.

5.2.1 **Overall Forecasting Performance (RQ1).** Table 2 reports the experimental results of all compared methods in forecasting traffic accidents over different time periods in 2017 and 2018. We can observe that *DFN* always achieves the best performance in all cases, which ascertains the effectiveness of the proposed approach. Notably, the performance gains over various strong baselines is large in terms of *F1-score*, *AUC* and *AP*. The advance of *DFN* over all baselines indicates that *DFN* is able to effectively transfer knowledge from both static and dynamic external factors in assisting the traffic accident forecasting task.

The performance is followed by ASM, which leverages attention network to endow the representation capability on spatial-temporal sequences. This further demonstrates the utility of augmenting the recurrent neural network in an interpret way. However, ASM ignores the hierarchically structured correlations between other urban data sources, which could easily lead to suboptimal spatial-temporal representations. Compared to S-LSTM and ST-RNN, the performance improvement might be attributed to the temporal aggregation mechanism–automatically specifying the attentive weights of each learned spatial-temporal pattern embedding.

In addition, *DFN* consistently achieves better performance than CFPT and Wide&Deep. It demonstrates the necessarity of modeling both traffic accidents and external heterogeneous spatial-temporal data in a dynamic manner. While CFPT projects region-time relations into latent vectors, it fails to capture the evolving temporal patterns of temporally ordered accident sequences.

5.2.2 **Forecasting Performance v.s. Time Windows (RQ2).** To investigate the forecasting performance over different time windows (reflective of various seasonal periods), we aim to predict traffic accidents by varying training/validation/test periods with sliding time windows, *i.e.*, Winter: Dec-Feb; Spring: Mar-May; Summer: Jun-Aug; Fall: Sep-Nov. As we have observed, *DFN* consistently yields best performance when competing with other baselines. Such remarkable improvement indicates the robustness of *DFN* in forecasting traffic accidents over different time periods.

5.3 Effect of Heterogeneous Factor Aggregation in *DFN* (RQ3)

To investigate whether exploiting heterogeneous factors helps to achieve better prediction accuracy, we perform experiments to show the effect of incorporating different external factors in our *DFN* with two variants: *DFN*-p and *DFN*-a represent the traffic accident prediction framework without aggregating POI and urban abnormal event data, respectively. We show the evaluation results in terms of F1-score, AUC and AP in Figure 3 and summarize the key findings as follows:

- The evaluation results indicate that semantic-aware region embedding generated from the designed context-embedding module could benefit the model performance. It demonstrates that mapping the latent spatial relations across regions into on semantic space is important to model complex patterns of traffic accident occurrences.
- Without the aggregation of urban complaint data, the learned spatial-temporal representation for each traffic accident series depends on only the static POI information. The performance gain between our defualt version *DFN* and *DFN*-a justifies that: exploiting dynamic high-order inter-dependencies between target traffic accident data and urban abnormal events greatly facilitates the representation capacity of our learning system on time-sensitive spatial-temporal data.

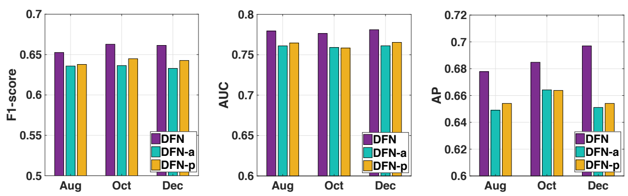
Hence, the above observations verify that aggregation of heterogeneous external factor into our predictive solution is helpful to learn more complex transition patterns of traffic accident occurrences.

5.4 Model Architecture Ablation Study (RQ4)

Additionally, we perform ablation experiments over a number of key components of *DFN* in order to better understand their impacts,

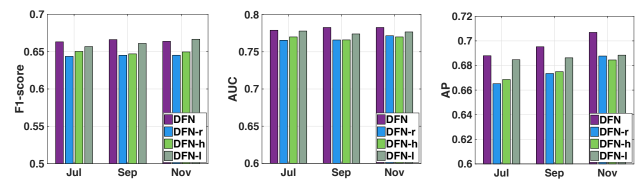
Table 2: Performance comparisons of all methods in predicting traffic accidents.

Period	2017-Jul			2017-Aug			2017-Sep			2017-Oct			2017-Nov			2017-Dec		
Method	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP
LR	0.546	0.645	0.518	0.531	0.637	0.498	0.557	0.601	0.465	0.546	0.637	0.513	0.568	0.565	0.453	0.581	0.654	0.527
W&D	0.577	0.753	0.647	0.588	0.729	0.616	0.601	0.730	0.630	0.579	0.731	0.632	0.593	0.732	0.630	0.581	0.728	0.614
SVR	0.565	0.735	0.621	0.574	0.735	0.609	0.577	0.737	0.628	0.567	0.732	0.613	0.578	0.739	0.642	0.573	0.737	0.626
ARIMA	0.559	0.760	0.657	0.563	0.758	0.644	0.566	0.762	0.668	0.557	0.754	0.653	0.573	0.767	0.677	0.564	0.760	0.667
S-LSTM	0.587	0.741	0.594	0.582	0.739	0.624	0.594	0.752	0.656	0.576	0.747	0.653	0.619	0.757	0.665	0.609	0.751	0.654
ST-RNN	0.574	0.754	0.648	0.575	0.749	0.634	0.583	0.753	0.657	0.571	0.748	0.655	0.587	0.759	0.668	0.581	0.755	0.658
ASM	0.594	0.755	0.649	0.598	0.750	0.635	0.584	0.753	0.658	0.575	0.748	0.654	0.604	0.759	0.668	0.606	0.759	0.668
CFPT	0.591	0.774	0.679	0.584	0.768	0.661	0.593	0.766	0.677	0.605	0.765	0.673	0.586	0.769	0.686	0.581	0.764	0.672
DFN	**0.663**	**0.779**	**0.688**	**0.653**	**0.779**	**0.678**	**0.666**	**0.783**	**0.695**	**0.663**	**0.776**	**0.685**	**0.664**	**0.783**	**0.707**	**0.661**	**0.781**	**0.697**
Period	2018-Jan			2018-Feb			2018-Mar			2018-Apr			2018-May			2018-Jun		
LR	0.542	0.616	0.462	0.546	0.682	0.538	0.564	0.686	0.566	0.555	0.683	0.553	0.557	0.652	0.543	0.570	0.648	0.554
W&D	0.577	0.721	0.592	0.582	0.736	0.611	0.589	0.714	0.598	0.586	0.748	0.635	0.579	0.729	0.640	0.582	0.730	0.655
SVR	0.557	0.735	0.596	0.561	0.740	0.604	0.567	0.734	0.613	0.559	0.730	0.606	0.570	0.738	0.642	0.577	0.739	0.656
ARIMA	0.571	0.716	0.581	0.555	0.642	0.499	0.562	0.656	0.530	0.558	0.671	0.532	0.542	0.675	0.569	0.551	0.739	0.657
S-LSTM	0.568	0.741	0.615	0.597	0.751	0.611	0.581	0.751	0.639	0.588	0.746	0.630	0.595	0.753	0.668	0.604	0.758	0.681
ST-RNN	0.587	0.745	0.624	0.571	0.753	0.629	0.578	0.751	0.643	0.590	0.748	0.641	0.590	0.756	0.673	0.597	0.759	0.684
ASM	0.588	0.745	0.624	0.581	0.753	0.630	0.596	0.752	0.643	0.592	0.747	0.640	0.593	0.756	0.674	0.601	0.759	0.683
CFPT	0.590	0.760	0.627	0.584	0.764	0.651	0.588	0.769	0.670	0.588	0.742	0.592	0.586	0.761	0.671	0.616	0.768	0.693
DFN	**0.632**	**0.773**	**0.659**	**0.632**	**0.775**	**0.652**	**0.652**	**0.775**	**0.681**	**0.651**	**0.779**	**0.680**	**0.666**	**0.775**	**0.703**	**0.681**	**0.786**	**0.712**



(a) F1-score (b) AUC (c) AP

Figure 3: Effect of Heterogeneous Factor Aggregation.



(a) F1-score (b) AUC (c) AP

Figure 4: Model Architecture Ablation Study of *DFN*.

including hierarchical fusion networks, temporal aggregation mechanism and recurrent unit selection. Figure 4 shows the results of performance comparison between our designed default version *DFN* and other three model variants (corresponding to different module aspects):

- **Impact of Pattern Fusion Module**: *DFN*-h. A simplified version of *DFN* without the fusion layer between the recurrent neural networks for encoding traffic accident sequence and urban complaint sequence in chronological way.
- **Effect of Temporal Aggregation Mechanism**: *DFN*-r. Another simplified version of *DFN*, which directly utilizes the latent representations learned from hierarchical fusion networks to predict the future accident occurrence, *i.e.*, without employing the aggregation mechanism.
- **Effect of Recurrent Unit Selection**: *DFN*-l. It selects the LSTM as the recurrent unit in our hierarchical fusion networks to capture temporal patterns of sequences.

Based on the evaluation results, we analyze each module effects respectively: 1) The results show that removing hierarchical fusion networks causes *DFN*'s performance decreasing. The lack of stacking more recurrent layers for fusing inter-correlations between traffic accidents and dynamic urban report environment phenomena, makes the forecasting model ill-posed. 2) We can observe the integration of temporal aggregation layer can boost forecasting performance. This verifies that it is helpful to learn more comprehensive transition regularities of traffic accidents via attention architecture. 3) We could notice that the selection of recurrent unit in our hierarchical fusion networks has low impact on the model performance. We adopt GRU in our *DFN* since it incurs a smaller computational cost.

5.5 Parameter Effect Investigation (RQ5)

In this subsection, we seek to examine the effect of hyperparameter settings in our *DFN*. Figure 5 illustrates the impact of the embedding dimension e, sequence length T and attention size S in forecasting traffic accidents on May 2018. The evaluated model performance is measured in terms of F1-score, AUC and AP. We keep the default parameter settings when varying the value of the target hyperparameter.

Impact of embedding dimension e. We can observe that the forecasting performance becomes better as the embedding dimension e increases, since at the early stage. An interesting observation is that a larger hidden dimension does not necessarily leads to better model performance. This phenomenon might be attributed to the overfitting issue.

Impact of sequence length T. The parameter T controls the sequence length of the input traffic accident data. Observe from Figure 5, using the sequence length $T = 10$ achieves the best performance. In our experiments, setting T between 8 and 12 is more preferable. The above observation also reflects the efficacy of our temporal aggregation mechanism in augmenting sequential transition modeling with a relatively shorter input sequence.

Impact of attention size S. We can notice that the model performance becomes better as S increases and tends to saturate once $S = 256$. While a larger value of attention size S brings a stronger representation power for our attentive aggregation network, setting attention size S to a larger value may not be continually beneficial to the performance.

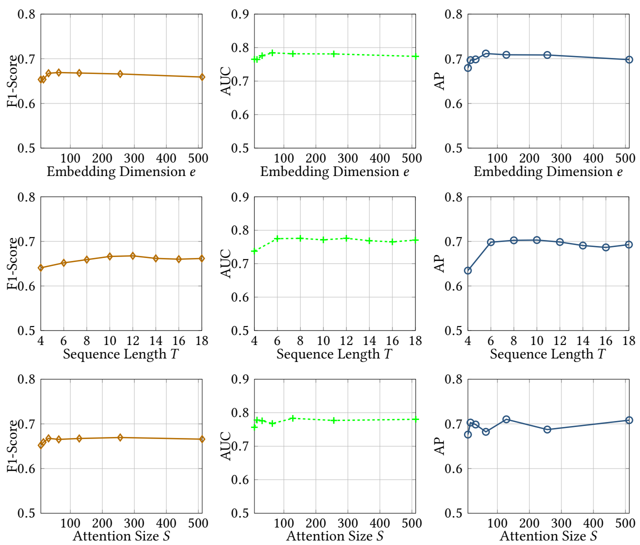
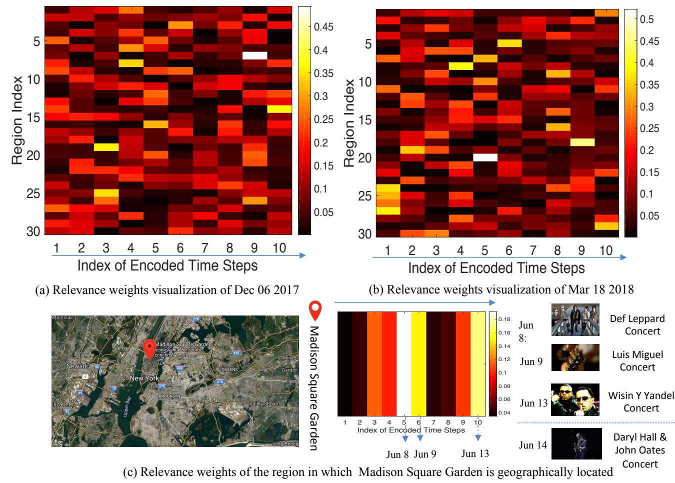


Figure 5: Hyperparameter Studies of *DFN* *w.r.t* embedding dimension e, attention size S and sequence length T.

5.6 Model Interpretation Study (RQ6)

In real-world scenarios, it is important to understand what contributes to the forecasting results. Hence, apart from the quantitative analysis of traffic accident forecasting results, we are interested in providing qualitative examples to better understand the aggregation mechanism in *DFN*. The visualization results of Figure 6(a) and (b) corresponds to the learned relevance scores in predicting traffic accidents at 30 sampled regions on Dec 06, 2017 (weekday) and Mar 18, 2018 (weekend), respectively. We can notice that *DFN* enables the dynamic modeling of correlations between the target time steps and previous encoded time steps. This better model interpretability is particularly important for policy makers.

Furthermore, we also take the region in which the Madison Square Garden is geographically located (see Figure6(c)) as an example, to visualize the relevance weights for all encoded time steps. In NYC, Madison Square Garden is a well-known multipurpose arena to hold events, such as concerts and games. We visualize the attention weights in predicting crimes on Jun 14, 2018 as shown in Figure 6(c). In this figure, an interesting observation is that the traffic accident occurrences on Jun 14 at this region is more relevant to that on Jun 13, Jun 9 and Jun 8, *i.e.*, with higher relevance scores with those three time slots. This is affected by the three entertainment events held on Jun, 8, 9 and 13, respectively (namely, Jun 8:



(a) Relevance weights visualization of Dec 06 2017 (b) Relevance weights visualization of Mar 18 2018

(c) Relevance weights of the region in which Madison Square Garden is geographically located

Figure 6: (a)-(b) The attention matrix obtained from the forecasting on Dec 06 2018 (weekday) and on Mar 18 2018 (weekend), respectively, where each row represents the attention vector of true positive region cases over the inputs. (c) The attention matrix obtained from traffic accident predictions at the region in which the Madison Square Garden is geographically located on Jun 14 2018.

Def Leppard concert, Jun 9: Luis Miguel concert, Jun 13: Wisin Y Yandel concert). Also, there is a concert (Daryl Hall & John Oates Concert) held on Jun 14. These observations suggest that our *DFN* framework is not only effective in forecasting accuracy but can also be easily interpreted in predicting future traffic accidents.

6 CONCLUSION AND FUTURE WORK

In this work, we develop a novel framework named DFN for better traffic accident forecasting. Particularly, DFN encodes the latent semantic signals into the region representations, and tracks the dynamic inter-dependencies between the traffic accident data and external factors of urban events by hierarchical fusion network. We further augment our predictive solution with a temporal-wise recalibration mechanism to learn complex sequential transitions of traffic accident occurrences. The evaluation on real-world transportation datasets to show that our model achieves the best performance against various baselines.

One future direction deserves more investigation is how to extend our DFN to a recursive framework, in order to handle streaming traffic accident records arriving continually. It is interesting to design a recursive algorithm over a batch model architecture to enable more efficient online traffic accident forecasting. One potential solution is how to leverage model hyperparameters learned from previous training period for parameter updating at the current training phase. Furthermore, it is also interesting to deploy our framework under a cloud/client architecture for collecting data in a timely manner and visualize prediction results.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint*

arXiv:1409.0473 (2014).

[2] Lida Barba, Nibaldo Rodríguez, and Cecilia Montt. 2014. Smoothing strategies combined with ARIMA and neural networks to improve the forecasting of traffic accidents. *The Scientific World Journal* 2014 (2014).

[3] Ruth Bergel-Hayat, Mohammed Debbarh, Constantinos Antoniou, and George Yannis. 2013. Explaining the road accident risk: weather effects. *Accident Analysis & Prevention* 60 (2013), 456–465.

[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *Transactions on Intelligent Systems and Technology (TIST)* 2, 3 (2011), 27.

[5] Heng-Tze Cheng, Levent Koc, et al. 2016. Wide & deep learning for recommender systems. In *Recommender Systems Conference (Recsys)*. ACM, 7–10.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, et al. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[7] Rupam Deb and Alan Wee-Chung Liew. 2016. Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences* 339 (2016), 274–289.

[8] Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V Chawla, and Bai Wang. 2015. Coupledlp: Link prediction in coupled networks. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 199–208.

[9] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *International Conference on World Wide Web (WWW)*. ACM, 1459–1468.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *International Conference on World Wide Web (WWW)*. ACM, 173–182.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression.* Vol. 398. John Wiley & Sons.

[13] Chao Huang, Dong Wang, and Shenglong Zhu. 2017. Where are you from: Home location profiling of crowd sensors from noisy and sparse crowdsourcing data. In *International Conference on Computer Communications (INFOCOM)*. IEEE, 1–9.

[14] Chao Huang, Xian Wu, and Dong Wang. 2016. Crowdsourcing-based urban anomaly prediction system for smart cities. In *International on Conference on Information and Knowledge Management (CIKM)*. ACM, 1969–1972.

[15] Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Nitesh Chawla, and Dawei Yin. 2019. MiST: A Multiview and Multimodal Spatial-Temporal Learning Framework for Citywide Abnormal Event Forecasting. In *The World Wide Web Conference (WWW)*. ACM, 717–728.

[16] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. 2018. DeepCrime: Attentive Hierarchical Recurrent Networks for Crime Prediction. In *International Conference on Information and Knowledge Management (CIKM)*. ACM, 1423–1432.

[17] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015).

[18] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *International Conference on Learning Representations (ICLR)* (2018).

[19] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[21] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. 2013. Predicting taxi–passenger demand using streaming data. *Transactions on Intelligent Transportation Systems (TITS)* 14, 3 (2013), 1393–1402.

[22] Randa Oqab Mujalli, Griselda López, and Laura Garach. 2016. Bayes classifiers for imbalanced traffic accidents datasets. *Accident Analysis & Prevention* 88 (2016), 37–51.

[23] VA Olutayo and AA Eludire. 2014. Traffic accident analysis using decision trees and neural networks. *International Journal of Information Technology and Computer Science* 2 (2014), 22–28.

[24] World Health Organization. 2015. *Global status report on road safety 2015.* World Health Organization.

[25] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 2627–2633.

[26] Honglei Ren, You Song, JingXin Liu, Yucheng Hu, and Jinzhi Lei. 2017. A Deep Learning Approach to the Prediction of Short-term Traffic Accident Risk. *arXiv preprint arXiv:1710.09543* (2017).

[27] Bharti Sharma, Vinod Kumar Katiyar, and Kranti Kumar. 2016. Traffic accident prediction model using support vector machines with Gaussian kernel. In *ICSC*. Springer, 1–10.

[28] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level.. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 16. 2618–2624.

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research (JMLR)* 15, 1 (2014), 1929–1958.

[30] JD Tamerius, X Zhou, R Mantilla, and T Greenfield-Huitt. 2016. Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions. *Weather, Climate, and Society* 8, 4 (2016), 399–407.

[31] Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. 2015. Extreme learning machine for multilayer perceptron. *Transactions on Neural Networks and Learning Systems (TNNLS)* 27, 4 (2015), 809–821.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, and etc. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*. 5998–6008.

[33] Chenglong Wang, Feijun Jiang, and Hongxia Yang. 2017. A hybrid framework for text modeling with convolutional rnn. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2061–2069.

[34] Jing Wang and Xin Wang. 2011. An ontology-based traffic accident risk mapping framework. In *Symposium on Spatial and Temporal Databases (SSTD)*. Springer, 21–38.

[35] Senzhang Wang, Xiaoming Zhang, Jianping Cao, Lifang He, Leon Stenneth, Philip S Yu, Zhoujun Li, and Zhiqiu Huang. 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. *Transactions on Information Systems (TOIS)* 35, 4 (2017), 40.

[36] Xian Wu, Yuxiao Dong, Chao Huang, Jian Xu, Dong Wang, and Nitesh V Chawla. 2017. Uapd: Predicting urban anomalies from spatial-temporal data. In *The European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. Springer, 622–638.

[37] Xian Wu, Baoxu Shi, Yuxiao Dong, Chao Huang, Louis Faust, and Nitesh V Chawla. 2018. RESTFul: Resolution-Aware Forecasting of Behavioral Time Series Data. In *International Conference on Information and Knowledge Management (CIKM)*. ACM, 1073–1082.

[38] Carl Yang, Lanxiao Bai, Chao Zhang, Quan Yuan, and Jiawei Han. 2017. Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In *International Conference on Knowledge Discovery & Data Mining (KDD)*. ACM, 1245–1254.

[39] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[40] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, and Jieping Ye. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *AAAI Conference on Artificial Intelligence (AAAI)*. 2588–2595.

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. 4651–4659.

[42] Rose Yu, Yaguang Li, Cyrus Shahabi, Ugur Demiryurek, and Yan Liu. 2017. Deep learning: A generic approach for extreme condition traffic forecasting. In *SIAM International Conference on Data Mining (SDM)*. SIAM, 777–785.

[43] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 186–194.

[44] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2014. Discovering urban functional zones using latent activity trajectories. *Transactions on Knowledge and Data Engineering (TKDE)* 27, 3 (2014), 712–725.

[45] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 984–992.

[46] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 271–278.

[47] Chao Zhang, Mengxiong Liu, Zhengchao Liu, Carl Yang, Luming Zhang, and Jiawei Han. 2018. Spatiotemporal Activity Modeling Under Data Scarcity: A Graph-Regularized Cross-Modal Embedding Approach. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[48] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Multi-resolution Spatial Event Forecasting in Social Media. In *International Conference on Data Mining (ICDM)*. IEEE, 689–698.

[49] Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. 2017. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *International Conference on World Wide Web (WWW)*. ACM, 153–162.