Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Special Track on AI for Computational Sustainability and Human Well-being

Cross-Interaction Hierarchical Attention Networks for Urban Anomaly Prediction

Chao Huang[1], **Chuxu Zhang**[2], **Peng Dai**[1], **Liefeng Bo**[1]
JD Finance America Corporation[1]
Brandeis University[2]
chaohuang75@gmail.com, chuxuzhang@brandeis.edu, {peng.dai, liefeng.bo}@jd.com

Abstract

Predicting anomalies (*e.g.*, blocked driveway and vehicle collisions) in urban space plays an important role in assisting governments and communities for building smart city applications, ranging from intelligent transportation to public safety. However, predicting urban anomalies is not trivial due to the following two factors: i) The sequential transition regularities of anomaly occurrences is complex, which exhibit with high-order and dynamic correlations. ii) The Interactions between region, time and anomaly category is multi-dimensional in real-world urban anomaly forecasting scenario. How to fuse multiple relations from spatial, temporal and categorical dimensions in the predictive framework remains a significant challenge. To address these two challenges, we propose a Cross-Interaction Hierarchical ATtention network model (CHAT) which uncovers the dynamic occurrence patterns of time-stamped urban anomaly data. Our CHAT framework could automatically capture the relevance of past anomaly occurrences across different time steps, and discriminates which types of cross-modal interactions are more important for making future predictions. Experiment results demonstrate the superiority of CHAT framework over state-of-the-art baselines.

1 Introduction

Urban anomalies has drawn increasing attention with the growing number of urban sensing platforms (*e.g.*, noise monitoring system and traffic condition reporting sites) [Zheng *et al.*, 2014; Wu *et al.*, 2020]. This work aims to predict anomalies of different categories at each region of a city to enable more timely and efficient resolution of urban issues. Predicting urban anomalies is of great value to traffic management and intelligent transportation. For example, if one can forecast blocked driveway events and traffic accidents beforehand, such anomalies can be prevented or mitigated by utilizing emergency mechanisms or designing more effective strategies (*e.g.*, traffic flow control [Iordanidou *et al.*, 2017]).

Existing work to anomaly prediction and detection in a city merely focus on summarizing temporal occurrence transitions linearly while cannot handle the scenarios where dynamic temporal dependencies exist in the distributions of anomalies [Wu *et al.*, 2017; Huang *et al.*, 2016; Zheng *et al.*, 2015]. As shown in [Zhang *et al.*, 2017; Zhang *et al.*, 2018], the time factor plays an important role in modeling urban anomalies, since regions, time periods and anomaly categories are dynamic correlated in urban spaces. Hence, failing to effectively capture such subtle dynamics is inappropriate and cannot represent the complicated real-world scenario. In contrast, this paper explicitly addresses the problem of urban anomaly prediction under dynamic scenarios.

However, developing such an urban anomaly prediction system is very difficult and two important technical challenges exist. *First*, the patterns of anomalies may vary over time, *e.g.*, anomaly causality in summer might be different from that in winter. Traditional time series models (*e.g.*, autoregression integrated moving average (ARIMA) [Wiesel *et al.*, 2013] and Gaussian Processing (GP) [Esling and Agon, 2012]) do not well capture the complex spatial-temporal dependencies in a fully dynamic manner [Liu *et al.*, 2016]. While recurrent neural network (RNN) and its variants (*e.g.*, LSTM and GRU) have been utilized in modeling non-linear transition regularities of spatial-temporal data [Yu *et al.*, 2017; Liu *et al.*, 2016], these models can hardly capture long-term temporal correlations from a global perspective, and understand which historical anomaly occurrences should be more emphasized during the prediction phase.

Another challenge is how to effectively model time-evolving multi-dimensional interactions. In real-world urban scenarios, interactions across regions (spatial dimension), time frames (temporal dimension) and anomaly categories (semantic dimension) are implicit and time-evolving [Feng *et al.*, 2018], which makes the urban anomaly prediction more challenging. For example, blocked driveway may have different probabilities of occurrence at different regions across different time slots, due to some non-periodic events (*e.g.*, short-term road constructions). Hence, cross-dimensional interactions at different time frames might have varying importance in helping the urban anomaly prediction task. Effectively capturing the interaction importance remains a challenge.

To tackle the aforementioned challenges, we propose a new deep learning framework–Cross-Interaction Hierarchical ATtention Network (CHAT). We first propose to encode the dynamic anomaly occurrence patterns by integrating the

temporal-wise attention mechanism and bidirectional long short-term memory network. The integrative framework augments recurrent neural architecture by learning an attentive gating mechanism, which is calculated based on the summarized global temporal information, to recalibrate learned time-ordered sequential transitional regularities from past observations across different time slots. In addition, to fully exploit the time-evolving multi-dimensional interactions in modeling urban anomaly data, we further develop an interaction-wise attention mechanism to learn the joint representations of anomaly occurrence patterns across spatial-temporal-semantic dimensions. Our developed CHAT model leverages the augmented recurrent layer to discover the temporal dependency patterns, and the strengths of attention mechanism to capture complex time-evolving multi-dimensional interactions across time, location and categories.

In summary, we highlight our contributions as follows:

- In this work, we explore the urban anomaly prediction problem from the viewpoint of hierarchical attention networks, empowering it to effectively model time-evolving multi-dimensional spatial-temporal data.
- We propose an integrative architecture with bi-directional recurrent layer and temporal-wise attention mechanism, to exploit global temporal contextual information.
- Additionally, we design an interaction-wise attention mechanism to learn the joint representations across spatial-temporal-semantic dimensions for consensus anomaly prediction in a seamless manner.
- We conduct extensive experiments on the real-world urban anomaly datasets to show that our developed CHAT framework consistently outperforms state-of-the-art methods.

2 Preliminaries

We consider a set of I geographic regions in a city: $R_1, ..., R_I$, a set of J anomaly categories: $O_1, ..., O_J$, and K time slots, where i, j and k is used to represent the index for region, anomaly category and time slot, respectively. Anomalies of each region are reported from the first time slot (e.g., day) to K-th time slot.

Definition 1 **Anomaly Tensor** $\mathcal{AS}$. We define an anomaly tensor $\mathcal{AS}_i \in \mathbb{R}^{I \times J \times K}$ to represent the time-stamped anomaly sequences of all geographical regions for all anomaly categories across K time slots. In particular, we set the element $\mathcal{AS}_{i,j,k} = 1$ given the O_j-th category of anomalies reported from region the region R_i at the k-th time slot. Otherwise, the $\mathcal{AS}_{i,j,k}$ will be set as 0.

Problem Statement. Based on the above defined terms, we formulate the urban anomaly prediction problem as follows: given the anomaly tensor $\mathcal{AS}$ generated from the first time slot till K-th time slot, the goal is to forecast the anomaly occurrence of each category O_j at each region R_i of a city in the future time slot.

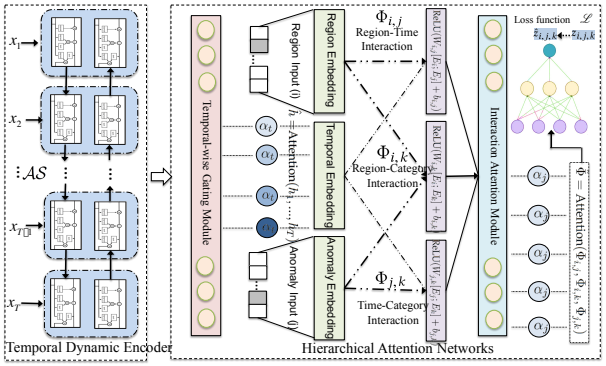


Figure 1: The Architecture of CHAT Model.

3 Methodology

3.1 Dynamic Temporal Dependencies Modeling

While Recurrent Neural Networks (RNNs) have been widely applied to model sequential data by proposing hidden-to-hidden connections [Bishop, 1995], they are unable to learn the long-term dependencies (*i.e.*, dependencies between time steps that are far apart) in time series data. Hence, we utilize Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] to address the above limitations and make our predictive model more effective. Formally, the vector representations of hidden states h_t and c_t for each time step t are derived with the following operations:

$$\begin{aligned} i_t &= \sigma(W_i \cdot h_{t-1} + V_i \cdot x_t + b_i) \\ o_t &= \sigma(W_o \cdot h_{t-1} + V_o \cdot x_t + b_o) \\ f_t &= \sigma(W_f \cdot h_{t-1} + V_f \cdot x_t + b_f) \\ \widetilde{c_t} &= \phi(W_c \cdot h_{t-1} + V_c \cdot x_t + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \tag{1}$$

where $W_* \in \mathbb{R}^{d_s \times d_s}$ represents the learnable weight matrix for the representation c_{t-1} and h_{t-1} from the $t-1$ time slot. $V_* \in \mathbb{R}^{d_x \times d_s}$ is another weight matrix for the input. b_* is the defined bias term. The latent dimensionality of input and latent embedding vector is respectively denoted by d_x and d_s. The activation function of tanh ($\phi(\cdot)$) and sigmoid ($\sigma(\cdot)$) function is applied over the transformation operation. In our recurrent unit, the input input gate, output gate and forget gate is defined as i_t, o_t, and f_t, respectively. We represent the operations in the LSTM unit as $[c_t, h_t] = \text{LSTM}(*, c_{t-1}, h_{t-1})$.

Furthermore, to improve the performance of LSTM in modeling long sequence data, Bidirectional Long Short-Term Memory (BiLSTM) [Schuster and Paliwal, 1997] was developed to consider both the past and future contextual signals for each time step in the generated sequence. Specifically, BiLSTM involves two hidden layers separately with forward states $[\overrightarrow{c}_t, \overrightarrow{h}_t]$ and backward states $[\overleftarrow{c}_t, \overleftarrow{h}_t]$, respectively, *i.e.*, (1) forward hidden layer models the contextual information in sequence from 1-th to t-th time step, and (2) backward hidden layer models the contextual information in sequence from t-th to 1-th time step. In particular, we use the anomaly sequence from x_1 to x_T as the input to the forward LSTM and derive the sequence of forward hidden states $(\overrightarrow{h}_1, ..., \overrightarrow{h}_T)$

and $(\overrightarrow{c}_1, ..., \overrightarrow{c}_T)$. The backward LSTM takes the anomaly sequence information in the reverse order (*i.e.*, from x_T to x_1) as input to update the hidden states $(\overleftarrow{h}_1, ..., \overleftarrow{h}_T)$ and $(\overleftarrow{c}_1, ..., \overleftarrow{c}_T)$. We further concatenate the forward hidden state $\overrightarrow{h}_t$, $\overrightarrow{c}_t$ and backward hidden state $\overleftarrow{h}_t$, $\overleftarrow{c}_t$. The equations below describe how the forward and backward hidden states from individual time slot are updated.

$$\begin{aligned}[\overrightarrow{c_t}, \overrightarrow{h_t}] =& \text{LSTM}(x_t, \overrightarrow{c_{t-1}}, \overrightarrow{h_{t-1}}) \\ [\overleftarrow{c_t}, \overleftarrow{h_t}] =& \text{LSTM}(x_t, \overleftarrow{c_{t+1}}, \overleftarrow{h_{t+1}}) \\ h_t =& [\overrightarrow{h}_t; \overleftarrow{h}_t]^T \end{aligned} \tag{2}$$

Finally, we obtain the final hidden vector representation as $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]^T$.

3.2 Hierarchical Attention Networks

In order to mitigate the limitation of recurrent neural architecture in dealing with long-term dependencies [Graves, 2013], the attention mechanism was introduced to endow the neural network models with the capability of learning where to pay attention on the input series data, and generate the latent representations by differentiating the relevance of encoded time steps [Yang *et al.*, 2016]. The attention network introduces a context vector to learn different relationships between instances in an explicit manner.

Our hierarchical attention framework consists of two attention modules: (i) the temporal-wise gating module which aims to select the relevant anomaly information for predicting anomaly occurrence in the future; and (ii) the interaction attention which learns to score the importance of pair-wise interactions between regions, anomaly categories and time slots. In the following subsections, we will present the technical details of each component.

Temporal-wise Attention

The objective of anomaly prediction task is to forecast the occurrence of anomalies at each geographical region $R_i \in [R_1, ..., R_I]$ of a city in the future time slot, based on the observed anomaly data from previous time slots, *i.e.*, x_1 to x_T. To augment our bidirectional recurrent architecture with the ability in encoding long-term dependencies (with the large sequence length), and allow our CHAT framework to focus on certain parts of anomaly sequence, we design a temporal-wise attention layer and integrate it with the BiLSTM architecture to learn importance scores from a set of source hidden states.

We use $h_t, t \in [1, ..., T]$ to represent hidden state of t-th time slot. Then, we feed h_t into a one-layer MLP to obtain contextual vector u_t which corresponds to the hidden representation of h_t. A softmax function is further applied to calculate the relevance of each source hidden representation with a normalization operation. We formally model the relations between hidden states corresponding to different time slots using the temporal-wise attention mechanism as follows:

$$\begin{aligned} u_t =& tanh(W_{tem} h_t + b_{tem}) \\ \alpha_t =& \frac{exp(u_t^T u_{tem})}{\sum_t exp(u_t^T u_{tem})}; \quad \hat{h}_t = \sum_{t=1}^{T} \alpha_t h_t \end{aligned} \tag{3}$$

where $\hat{h}$ is defined as the concatenated hidden representation with the learned attention weights. Formally, we denote our temporal attention mechanism as below:

$$\hat{h} = \text{Attention}(h_1, ..., h_T). \tag{4}$$

Interaction Attention Module

With the consideration of spatial, temporal and categorical dimensions, we can represent the multi-dimensional anomaly data with a three-way tensor $\mathcal{AS}$, where each dimension stands for geographical region, anomaly category and time slot, respectively. Each element in this tensor indicates that: in k-th time slot, the anomaly occurrence of category O_j was reported from region R_i. Given this constructed tensor, we can note that the interactions exist between each pair of two individual dimensions [Rendle and Schmidt-Thieme, 2010] (*e.g.*, j-th category anomalies happened at region R_i or j-th category anomalies were reported from k-th time slot). In this work, we define E_i, E_j and E_k to represent the embedding vectors of region R_i, anomaly category O_j and k-th time slot, respectively. Here, we define the interactions between regions and anomaly categories, regions and time slots, anomaly categories and time slots as $\Phi_{i,j}$, $\Phi_{i,k}$, $\Phi_{j,k}$, respectively. The interaction between each two dimensions are formally defined as follows:

$$\begin{aligned} \Phi_{i,j} =& \text{ReLU}(W_{i,j}[E_i; E_j] + b_{i,j}) \\ \Phi_{i,k} =& \text{ReLU}(W_{i,k}[E_i; E_k] + b_{i,k}) \\ \Phi_{j,k} =& \text{ReLU}(W_{j,k}[E_j; E_k] + b_{j,k}) \end{aligned} \tag{5}$$

where W_* and b_* represents the learnable transmission matrix and bias term, respectively.

While the anomaly sequence AS of a entire city exhibits interactions between regions, anomaly categories and time slots, not all these three interactions contribute equally to help forecast future anomaly occurrence. Hence, we propose to utilize attention mechanism with an interaction level context vector ω which measures the importance of interactions between different dimensions. Our interaction attention can be given as follows:

$$\hat{\Phi} = \text{Attention}(\Phi_{i,j}, \Phi_{i,k}, \Phi_{j,k}) \tag{6}$$

We further denote the concatenated interaction as $\hat{\Phi}$ in our interaction attention as follows:

$$\begin{aligned} \omega =& tanh(W_{int}[\Phi_{i,j}, \Phi_{i,k}, \Phi_{j,k}] + b_{int}) \\ \alpha_j =& \frac{exp(u_j^T u_s)}{\sum_j exp(u_j^T u_s)}; \quad \sum_{j=1}^{t} = \alpha_j h_j \end{aligned} \tag{7}$$

3.3 Model Optimization of CHAT Model

In the learning process of our CHAT model, we aim to obtain the occurrence likelihood (denoted as $z_{i,j,k}$) of j-th category of anomalies at the k-th time slot from region R_i. In the loss function, the cross entropy is leveraged as follows:

$$\begin{aligned} \mathscr{L} =& - \sum_{(i,j,k) \in S} z_{i,j,k} \ln \hat{z}_{i,j,k} \\ & + (1 - z_{i,j,k}) \ln (1 - \hat{z}_{i,j,k}) \end{aligned} \tag{8}$$

The model parameters can be obtained by minimizing the defined loss function.

4 Evaluation

We perform extensive experiments to evaluate the effectiveness of our urban anomaly predictive framework (*i.e.*, *CHAT*) on the real-world dataset from New York City. In this section, we first describe the experimented data used in this work and introduce the detailed experimental settings. Then, we present the performance validation results to demonstrate the superiority of our developed framework with the comparison against state-of-the-art techniques. To be more specific, we aim to answer the following research questions:

- **RQ1**: Can our *CHAT* outperform state-of-the-art baselines in predicting urban anomalies over different time periods?
- **RQ2**: What is the performance of *CHAT* in predicting category-specific urban anomalies?
- **RQ3**: How does our *CHAT* perform in forecasting urban anomalies *w.r.t* different geographical region resolutions?
- **RQ4**: What is the impact of temporal-wise gating mechanism and interaction-wise attention mechanism in the joint deep neural network architecture?
- **RQ5**: How do hyperparameters affect the performance?

4.1 Experimental Setup

Data Description

We carry out experiments on the real-world urban anomaly dataset which is collected from New York City (NYC)[1]. This data contains different categories of urban anomaly reports from the 311 online platforms[2]. Each reported urban anomaly report is formatted as (timestamp information, anomaly category, coordinates). This dataset span from Jan 2014 to Dec 2014. We focus on four key categories of citywide anomalies (*e.g.*, Blocked Driveway, Building Use, Noise and Illegal Parking) in this work. Figure 2 shows the geographical distributions of different categories of anomalies in New York City (NYC) on August and October, respectively.

Region Partition

We map each anomaly report into one region of a city with different geographical scales, so as to investigate the performance of *CHAT* with respect to different region scales. We present the details of our partitioning methods as follows:

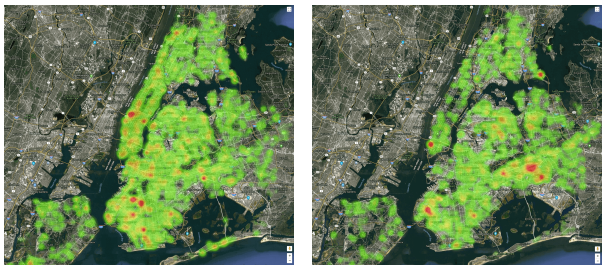
High-Level Geographical Region Scale. We partition New York City into 77 geographical areas based on the information of political districts [3]. Each individual partitioned geographical area is referred to as *high-level region*.

Fine-Grained Geographical Region Scale: We partition New York City into 862 geographical area [Zheng *et al.*, 2015] using road segments (*i.e.*, road segments from level 1 to level 5). Each individual partitioned geographical area is referred to as a *fine-grained region*.

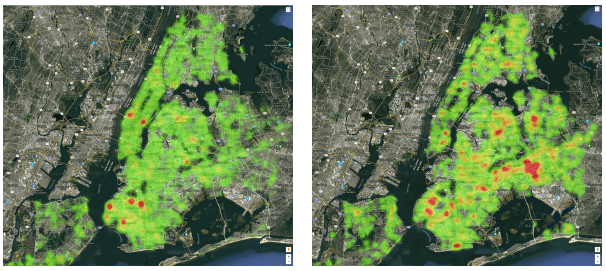
[1]https://data.cityofnewyork.us/

[2]https://portal.311.nyc.gov/

[3]https://data.cityofnewyork.us/Public-Safety/Police-Precincts/



(a) Illegal Parking-Aug (b) Building Use-Aug



(c) Illegal Parking-Oct (d) Building Use-Oct

Figure 2: Geographical distribution of anomaly occurrences across different categories on Aug and Oct.

Evaluation Protocols

We adopt two sets of evaluation metrics: we adopt *Marco-F1* and *Micro-F1* [Ni *et al.*, 2018] as the evaluation metrics to measure the prediction accuracy across different anomaly categories. We further use *F1-score* to investigate the performance of predicting category-specific anomalies. Note that higher Macro-F1, Micro-F1 and F1-score indicate better prediction performance.

In our experiments, the evaluation dataset is divided into into training, validation and test sets with the period of 5.5 month, 0.5 month and 0.5 month, respectively. The prediction results over all time slots (days) in the test period are averaged to generate the final prediction performance. During the evaluation process, the valuation set serve as the data for parameter tuning.

Baselines for Comparison

We consider six types of baselines for performance comparision: (i) traditional predictive analytic model (*i.e.*, LR); (ii) conventional time series forecasting techniques (*i.e.*, GP and ARIMA); (iii) Bayesian inference method for spatial-temporal data forecasting (*i.e.*, BIAP); (iv) tenor factorization technique for urban anomaly prediction (*i.e.*, TFAP). (v) deep recurrent networks for spatial-temporal data prediction (*i.e.*, ST-RNN and DRN); (vi) spatial-temporal pattern modeling with attentive recurrent framework (*i.e.*, ARF).

- **Gaussian Processing (GP)** [Esling and Agon, 2012]: it predicts the anomaly of each region using kernel function to measure the distance between past anomaly distribution and the anomaly occurrence in the predicted time slot.
- **Auto-Regressive Integrated Moving Average (ARIMA)** [Wiesel *et al.*, 2013]: this conventional time series approach aims to predict region's anomalies by

considering the frequency of anomaly occurrence.

- **Logistic Regression (LR)** [Hosmer Jr *et al.*, 2013]: we leverage this method with the incorporation of extracted temporal features from historical anomaly traces.
- **Tensor Factorization-based Anomaly Prediction (TFAP)** [Wu *et al.*, 2017]: It aims to predict anomaly occurrences by extending the Matrix Factorization scheme to consider the temporal dimension of crime data.
- **Bayesian Inference Anomaly Prediction (BIAP)** [Huang *et al.*, 2016]: it proposes a Bayesian inference method to jointly model the sequential patterns of anomalies and dependencies between different regions.
- **Spatial-Temporal Recurrent Neural Networks (ST-RNN)** [Liu *et al.*, 2016]: it proposes to model the spatial-temporal patterns of sequential data by employing the recurrent neural networks.
- **Deep Recurrent Networks (DRN)** [Yu *et al.*, 2017]: it is a stacked deep recurrent neural architecture to capture dynamic periodical transitional regularities for spatial-temporal data forecasting.
- **Attentive Recurrent Framework (ARF)** [Feng *et al.*, 2018]: this approach models the evolving dependencies in time-ordered spatial-temporal data with the integration of attention mechanisms and recurrent neural network.

Parameter Settings

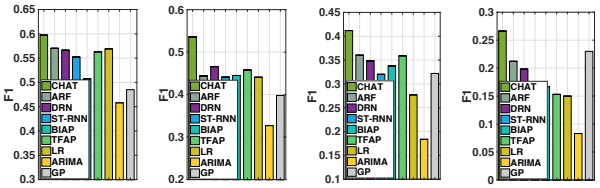
We use Adam [Kingma and Ba, 2015] as our optimizer to learn the model parameters of our *CHAT* framework. In our experiments, we set the hidden state dimensionality d and embedding dimension e as 32. Furthermore, the sequence length T in our recurrent neural architecture is set to 10. In the prediction phase of *CHAT*, we set the number of hidden layers as 3. The representation dimensionality of our attention mechanism is set to 32. The methods are trained from scratch without any pre-training on a single NVIDIA GeForce GTX 1080 Ti GPU with a learning rate and batch size of $1e^{-3}$ and 64.

4.2 Performance Comparison (RQ1, RQ2, RQ3)

Table 1 lists the evaluation results of all compared methods with different settings of training and test periods for fine-grained and high-level region scale, respectively. We observe that *CHAT* outperforms other baselines over different time periods. In addition, although different time windows reflect a spectrum of temporal diversity which is maintained by month and season variation (*e.g.*, Jul, Aug-Summer and Sep, Oct-Autumn), our proposed *CHAT* method consistently achieves the best performance by capturing such subtle temporal dynamics of urban anomaly data.

Note that the prediction task becomes more challenging when we map each urban anomaly into a specific fine-grained region (out of 862) compared to high-level region (out of 77), since the generated anomaly tensor AS will become more imbalanced by including more zero values, *i.e.*, there are fewer anomaly occurrences when mapping anomaly reports into more fine-grained geographical regions. we can observe that obvious improvements can be obtained by our *CHAT* with different geographical region scale as compared to competing baselines, suggesting that *CHAT* is capable of handing sparse urban anomaly data by explicitly jointly embedding all spatial, temporal, and categorical signals into the prediction.

We further perform experiments to evaluate *CHAT* in predicting individual anomaly categories with fine-grained region scale as shown in Figure 3. Overall, our proposed predictive system outperforms state-of-the-art methods in most cases. Additionally, *CHAT* achieves relatively 39.1%, 25.5% and 20.3% gain in terms of F1-score over ST-RNN, DRN and ARF respectively when predicting building use category with fine-grained geographical region scale. The advantage of the proposed *CHAT* lies in its proper consideration and accommodation of dynamic anomaly pattern challenge.



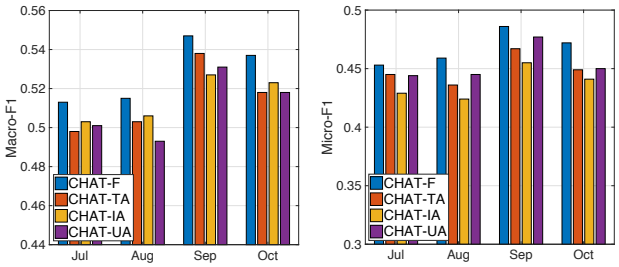
(a) Noise (b) Blocked Drive (c) Illegal Parking (d) Building Use

Figure 3: Forecasting results for category-specific anomalies.

4.3 Model Ablation Study (RQ4)

We now perform the model ablation study of the proposed *CHAT* with respect to the designed components. In this subsection, we consider three variants of the proposed *CHAT* method which correspond to different analytical aspects:

- **Effectiveness of Temporal Attention** *CHAT-IA*: A model variant without temporal-wise gating mechanism.
- **Effectiveness of Interaction Attention** *CHAT-TA*: Another variant of *CHAT* without interaction attention mechanism to capture cross-interaction patterns.
- **Effectiveness of Bi-directional Recurrent Architecture** *CHAT-UA*: This simplified version of *CHAT* models the anomaly sequence via unidirectional LSTM networks and dual-stage attention mechanisms.



(a) Macro-F1 (b) Micro-F1

Figure 4: Model Ablation Study of *CHAT* Framework.

Figure 4 presents the results of all compared variants in predicting anomalies across different categories respectively (with fine-grained region scale). We can notice that the full version of our framework *CHAT-F* achieves the best performance. We summarize the following key observations:

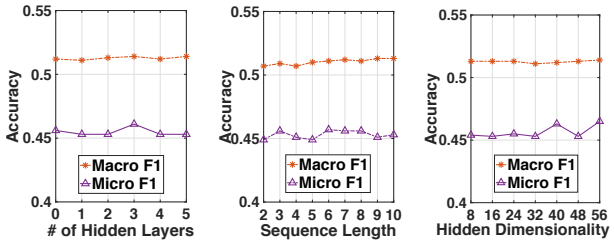
Scale	Fine-Grained Region						High-Level Region					
Month	July		August		September		July		August		September	
Method	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1	Mac-F1	Mic-F1
GP	0.391	0.359	0.402	0.364	0.423	0.385	0.773	0.756	0.789	0.769	0.795	0.773
ARIMA	0.332	0.263	0.342	0.261	0.360	0.282	0.818	0.782	0.826	0.788	0.834	0.793
LR	0.443	0.359	0.437	0.360	0.427	0.350	0.799	0.781	0.801	0.785	0.805	0.787
TFAP	0.477	0.398	0.485	0.385	0.493	0.421	0.833	0.818	0.850	0.834	0.849	0.832
BIAP	0.450	0.342	0.449	0.372	0.418	0.394	0.802	0.797	0.831	0.817	0.818	0.783
ST-RNN	0.439	0.369	0.464	0.386	0.472	0.392	0.823	0.810	0.840	0.825	0.839	0.824
DRN	0.460	0.395	0.465	0.388	0.459	0.378	0.826	0.813	0.832	0.817	0.834	0.820
ARF	0.468	0.404	0.478	0.404	0.478	0.399	0.831	0.813	0.829	0.815	0.838	0.823
CHAT	**0.513**	**0.453**	**0.515**	**0.459**	**0.547**	**0.486**	**0.867**	**0.841**	**0.875**	**0.854**	**0.882**	**0.861**

Table 1: Overall prediction results with both fine-grained and high-level geographical region scales across different periods.

- The efficacy of our designed temporal attention mechanism to encode the unknown relevance of past anomaly occurrences in forecasting future anomalies.
- The effectiveness of our interaction attention mechanism in capturing time-evolving multi-dimensional interactions across regions, time slots and anomaly categories.
- The positive effect of *CHAT* in modeling dynamic temporal dependencies with Bidirectional LSTM.

4.4 Parameter Study (RQ5)

We present the evaluation results of parameter study in *CHAT* in Figure 5. From the results, we summarize the following observations. *First*, we can see that the model performance tends to saturate once the embedding size reaches around 48. *Second*, we can observe that increasing the sequence length slightly improves overall performance. *Third*, we can observe the low impact of the number of hidden layers in the feed-forward network prediction layer on the performance.



(a) Hidden Layers (b) Sequence Length (c) Dimensionality

Figure 5: Parameter Study of *CHAT* Model.

5 Related Work

Urban Anomaly Detection and Forecasting. There exists a good amount of work on the topics *anomaly detection* [Pan *et al.*, 2013; Doan *et al.*, 2015; Zheng *et al.*, 2015; Le *et al.*, 2013]. For example, Doan *et al*. detected anomalies by modelling the behaviours of pedestrian flows across multiple locations [Doan *et al.*, 2015]. Zheng *et al*. identified the anomalies from spatial-temporal data using a probability-based detection method [Zheng *et al.*, 2015]. Nevertheless, the above schemes aim to identify the urban anomalies after they happen, which might lead to the inefficiency to handle the anomalies beforehand. Instead, this work tackles the problem of predict the urban anomalies before they happen. While there exist two recent work on anomaly prediction [Huang *et al.*, 2016; Wu *et al.*, 2017], significant limitations exist in their solutions: (i) they cannot handle the scenarios where dynamic temporal dependencies exist in the distributions of anomalies. (ii) They modeled the time-ordered anomaly sequence without differentiating the importance of past anomaly occurrences. To address those limitations, this work develops a new end-to-end prediction framework with the aim of modeling interactions between different dimensions from the urban anomaly data.

Time-stamped Data Modeling. Our work is related to the literature that focus on modeling time-stamped data [Shuai *et al.*, 2017; Huang *et al.*, 2018; Liu *et al.*, 2016; Hu *et al.*, 2017; Huang *et al.*, 2019; Zhang *et al.*, 2019]. Recent research efforts focus on applying recurrent neural network architectures for sequence modeling. Some example architectures include text parsing [Xiao *et al.*, 2017], scene segmentation [Shuai *et al.*, 2017] and location prediction [Huang *et al.*, 2017] and event forecasting [Hu *et al.*, 2017]. Different from the proposed methods, we present a new hierarchical attention-based neural network architecture to capture the dynamic patterns across time slots in chronological anomaly sequences and implicit interactions in multi-dimensional urban anomaly data.

6 Conclusion

In this work, we explored the neural network architectures to study the urban anomaly prediction problem by developing a new framework Cross-Interaction Hierarchical Attention Network (CHAT), to explicitly model the relation structures corresponding to different perspectives. Particularly, we first propose to capture the time-evolving dependencies in anomaly sequence with a bidirectional recurrent framework, so as to incorporate spatial and temporal context signals to enrich latent feature representations. Then, we design our method to carefully account for relations between regions, categories and time slots. We evaluate our new method on the real-world spatial-temporal dataset. The results showed that our scheme outperforms state-of-the-art baselines from different aspects. In future, a time-aware prediction model is needed to better handle streaming anomaly data and infer model parameters in a timely manner. In addition, it is also interesting to apply our CHAT method to other spatial-temporal learning applications with various urban urban sensing data.

References

[Bishop, 1995] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[Doan *et al.*, 2015] Minh Tuan Doan, Sutharshan Rajasegarar, et al. Profiling pedestrian distribution and anomaly detection in a dynamic environment. In *CIKM*, pages 1827–1830. ACM, 2015.

[Esling and Agon, 2012] Philippe Esling and Carlos Agon. Time-series data mining. *CSUR*, page 12, 2012.

[Feng *et al.*, 2018] Jie Feng, Yong Li, and etc. Deepmove: Predicting human mobility with attentional recurrent networks. In *WWW*, pages 1459–1468. ACM, 2018.

[Graves, 2013] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint:1308.0850*, 2013.

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

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

[Hu *et al.*, 2017] Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. What happens next? future subevent prediction using contextual hierarchical lstm. In *AAAI*, pages 3450–3456, 2017.

[Huang *et al.*, 2016] Chao Huang, Xian Wu, and Dong Wang. Crowdsourcing-based urban anomaly prediction system for smart cities. In *CIKM*. ACM, 2016.

[Huang *et al.*, 2017] Chao Huang, Dong Wang, and Shenglong Zhu. Where are you from: Home location profiling of crowd sensors from noisy and sparse crowdsourcing data. In *INFOCOM*, pages 1–9. IEEE, 2017.

[Huang *et al.*, 2018] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *CIKM*, pages 1423–1432. ACM, 2018.

[Huang *et al.*, 2019] Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Dawei Yin, and Nitesh Chawla. Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *WWW*, pages 717–728, 2019.

[Iordanidou *et al.*, 2017] Georgia-Roumpini Iordanidou, Ioannis Papamichail, et al. Feedback-based integrated motorway traffic flow control with delay balancing. *TITS*, 18(9):2319–2329, 2017.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[Le *et al.*, 2013] Viet Duc Le, Hans Scholten, and Paul Havinga. Flead: Online frequency likelihood estimation anomaly detection for mobile sensing. In *Ubicomp*, pages 1159–1166. ACM, 2013.

[Liu *et al.*, 2016] Qiang Liu, Shu Wu, et al. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, pages 194–200, 2016.

[Ni *et al.*, 2018] Jingchao Ni, Shiyu Chang, Xiao Liu, Wei Cheng, et al. Co-regularized deep multi-network embedding. In *WWW*, pages 469–478. ACM, 2018.

[Pan *et al.*, 2013] Bei Pan, Yu Zheng, et al. Crowd sensing of traffic anomalies based on human mobility and social media. In *SIGSPATIAL*, pages 344–353. ACM, 2013.

[Rendle and Schmidt-Thieme, 2010] Steffen Rendle and Lars Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, pages 81–90. ACM, 2010.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *TSP*, 45(11):2673–2681, 1997.

[Shuai *et al.*, 2017] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *TPAMI*, 2017.

[Wiesel *et al.*, 2013] Ami Wiesel, Ofir Bibi, et al. Time varying autoregressive moving average models for covariance estimation. *TSP*, pages 2791–2801, 2013.

[Wu *et al.*, 2017] Xian Wu, Yuxiao Dong, et al. Uapd: Predicting urban anomalies from spatial-temporal data. In *ECML/PKDD*. Springer, 2017.

[Wu *et al.*, 2020] Xian Wu, Chao Huang, Chuxu Zhang, and Nitesh V Chawla. Hierarchically structured transformer networks for fine-grained spatial event forecasting. In *WWW*, pages 2320–2330, 2020.

[Xiao *et al.*, 2017] Chunyang Xiao, Marc Dymetman, and Claire Gardent. Symbolic priors for rnn-based semantic parsing. In *IJCAI*, pages 4186–4192. Association for Computational Linguistics, 2017.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, et al. Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489. ACL, 2016.

[Yu *et al.*, 2017] Rose Yu, Yaguang Li, Cyrus Shahabi, and Yan Liu. Deep learning: A generic approach for extreme condition traffic forecasting. In *SDM*, 2017.

[Zhang *et al.*, 2017] Chao Zhang, Keyang Zhang, Quan Yuan, et al. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *WWW*, pages 361–370. ACM, 2017.

[Zhang *et al.*, 2018] Huichu Zhang, Yu Zheng, and Yong Yu. Detecting urban anomalies using multiple spatio-temporal data sources. *Ubicomp*, 2(1):54, 2018.

[Zhang *et al.*, 2019] Chuxu Zhang, Dongjin Song, Yuncong Chen, et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *AAAI*, volume 33, pages 1409–1416, 2019.

[Zheng *et al.*, 2014] Yu Zheng, Tong Liu, Yilun Wang, et al. Diagnosing new york city's noises with ubiquitous data. In *Ubicomp*, pages 715–725. ACM, 2014.

[Zheng *et al.*, 2015] Yu Zheng, Huichu Zhang, and Yong Yu. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *SIGSPATIAL*. ACM, 2015.