Foresee Urban Sparse Traffic Accidents: A Spatiotemporal Multi-Granularity Perspective

中国科学技术大学 计算机科学与技术学院

中国科大-数据智能实验室　　周正阳

2021.xx.xx

◆ 研究背景与价值意义

- 基于历史大数据提前预知不同场景下未来事故风险分布
- 提升**公共资源分配的精准性与公平性**
- 维护城市**公共安全**，提升幸福感和安全感

◆ 相关工作

【长期预测】针对全州（全省）下一周中**每日事故总数**进行预测，形成事故风险图，但均为固定的空间尺度。

【短期预测】1h的时间粒度，且均为**单步预测**，不能感知到路网的短期变化，未解决**本质稀疏**带来的零膨胀问题。

◆ 空间多尺度与短期多步预测

- 不同级别交警部门的多样化需求
- 市民出行规划和城市管理预警

Captured information and traffic volume analysis of a camera

Citywide traffic surveillance and accident spots

研究背景与相关工作回顾

◆ 研究背景与价值意义

- 基于历史大数据提前预知不同场景下未来事故风险分布
- 提升**公共资源分配的精准性与公平性**
- 维护城市**公共安全**，提升幸福感和安全感

◆ 相关工作

【长期预测】针对全州（全省）下一周中**每日事故总数**进行预测，形成事故风险图，但均为固定的空间尺度。

【短期预测】1h的时间粒度，且均为**单步预测**，不能感知到路网的短期变化，未解决**本质稀疏**带来的零膨胀问题。

精准防控风险
提升出行质量

问题：时空多尺度交通事故预测：给定静态路网结构特征$\mathcal{S}$ 和历史的动态交通信息 $\mathcal{F}(\Delta t)\,(\Delta t = 1, 2, \cdots T)$，我们的任务是同时预测在未来$r$步空间上粗粒度和细粒度的交通事故风险，以及最可能发生事故风险的M个区域，即 $\mathcal{O}_C(\Delta t')$， $\mathcal{O}_F(\Delta t')$，和 $\mathcal{V}_M(\Delta t')$，其中 $\Delta t' = T+1, T+2, \cdots T+r$ 。

RiskSeq: 基于多源历史数据预测未来多粒度事故分布

数据预处理: PKDE& ST-DFM

- 空间多粒度网格划分与稀疏动态数据实时推断
- 面向零膨胀问题的稀疏事故数据变换

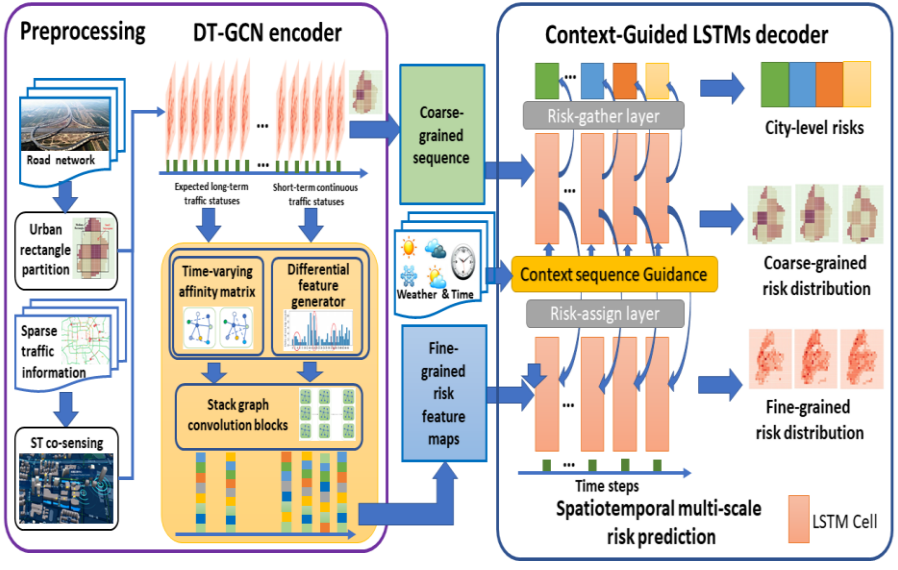
时空建模: DTGCN & CG-LSTM

- 路网动态变化与短期状态异常变化捕获
- 上下文引导的多步预测与空间多尺度依赖建模

后处理-事故筛选

- 自适应事故区域Ranking机制

Three-Phase Forecasting



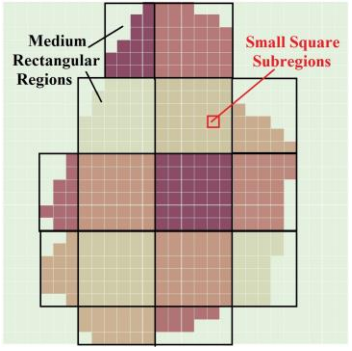
- 训练过程的零膨胀问题
- 有效数据覆盖面小，难以支撑训练

数据预处理

（1）层次性城市网格划分

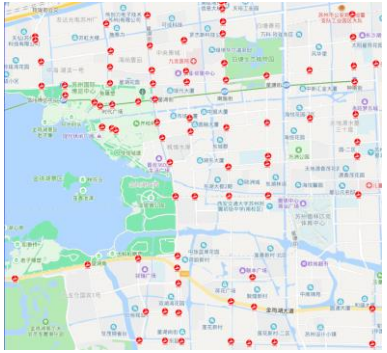
- ✓ 空间粒度降低，准确率提高
- ✓ 收集不同空间尺度的事故分布
- ✓ 便于设计事故区域提名策略

（2）本质稀疏与伪稀疏



（a）层次性城市划分

（b）本质稀疏


（c）伪稀疏

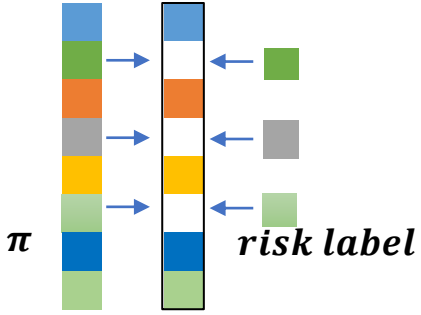
本质稀疏：本质稀疏就是数据记录本身偶发，即本身产生很少，即使全部获取也有限，如犯罪事件、交通拥堵（事故）事件、偶发疫情。无法补全。

伪稀疏： 数据随时随处产生，但是由于采集设备能力有限，无法全部获取（未完整感知）如城市路段通行速度 loop detector、气象/环境数据等。 可适当填充。

◆ 基于数据集先验信息的本质稀疏缓解方法-PKDE

基于先验信息的数据变换策略

- ✓ 扩大正负样本距离，显著区分潜在风险不同的区域；
- ✓ loss与label的一致性。



Step1: 统计每个区域在这个数据集上的事故总数，并归一化为0~1之间的概率值；

Step2: 将 ε_{v_i} 利用对数log转化成一个负数 π_{v_i}，并且使用一些约束的参数b1、b2来使其和正的risk风险值相一致，如正的风险值在0-5之间，那么负数值也在-5~0之间。

$$\varepsilon_{v_i} = \frac{1}{N_{week}} \sum_{j=1}^{N_{week}} \frac{r_{v_i}(j)}{\sum_{k=1}^{m} r_{v_k}(j)} \qquad \pi_{v_i} = b_1 \log_2 \varepsilon_{v_i} + b_2$$

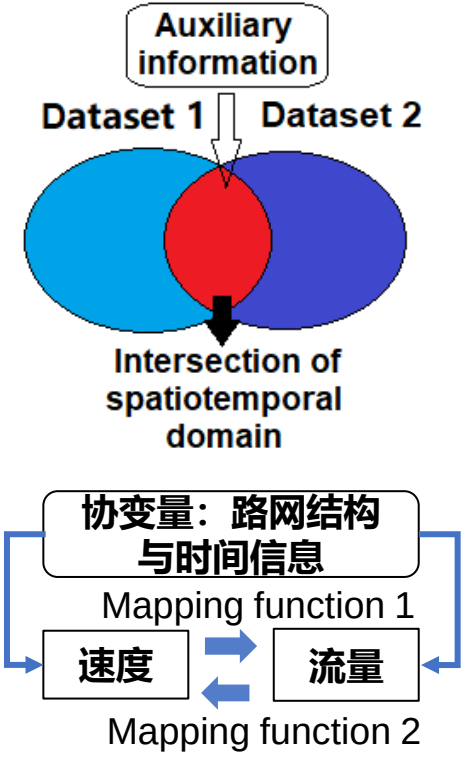
◆ 基于时空深度因子分解机的伪稀疏缓解策略 ST-DFM

基于数据集交叉时空域的协同推断学习

- 构造城市图 无向的城市图 $G(\mathcal{V},\mathcal{E})$. 其中$\mathcal{V}=\{v_1,v_2,v_i,\cdots,v_m\}$

$$\alpha_s(i,j)=\begin{cases}1 & \text{if subregion } v_i \text{ and } v_j \text{ are geographically adjacent}\\ e^{-JS(s_i\|s_j)} & \text{otherwise}\end{cases}$$

- 对每个区域而言，设计压缩交互网络CIN和DNN模块：
- 筛选对应的静态路网特征，时间戳分别放入1,2 field
- 筛选最邻近的区域的动态信息放入Real-time field



✓ 多源时空特征存在高阶交互影响，e.g. 路网结构与天气形成“共振”

✓ 车流与车速等动态特征存在非线性关联

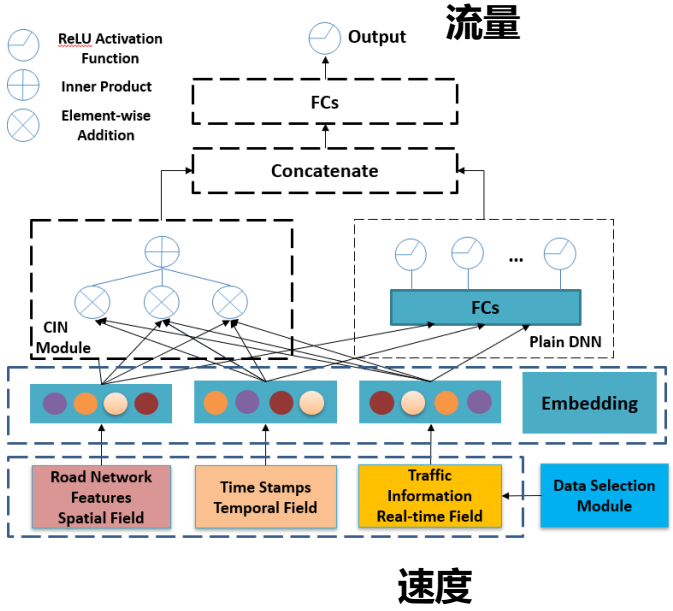
◆ 基于时空深度因子分解机的伪稀疏缓解策略 ST-DFM

基于数据集交叉时空域的协同推断学习

- 构造城市图　无向的城市图 $G(\mathcal{V},\mathcal{E})$. 其中$\mathcal{V}=\{v_1,v_2,v_i,\cdots,v_m\}$

$$\alpha_s(i,j)=\begin{cases}1 & \text{if subregion } v_i \text{ and } v_j \text{ are geographically adjacent}\\ e^{-JS(s_i\|s_j)} & \text{otherwise}\end{cases}$$

- 对每个区域而言，设计压缩交互网络CIN和DNN模块：
- 筛选对应的静态路网特征，时间戳分别放入1,2 field
- 筛选最邻近的区域的动态信息放入Real-time field



Motivation

✓ 多源时空特征存在高阶交互影响，
e.g. 路网结构与天气形成“共振”

✓ 车流与车速等动态特征存在非线性关联

时空建模：差分时变图卷积的神经网络 (Differential Time varying-GCN, DT-GCN)

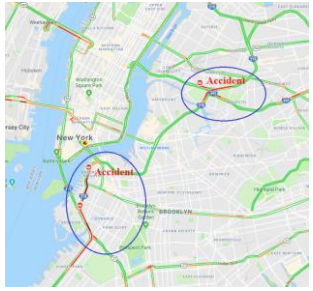
Why GCN:

（1）交通事故和道路拥堵存在一定的交互影响和传播关系；

（2）相似的路网结构和相似的动态交通模式可能事故共现；

（3）GCN适合建模传播关系与非欧氏关联。

两个处于交叉路口的
地区发生拥堵和事故



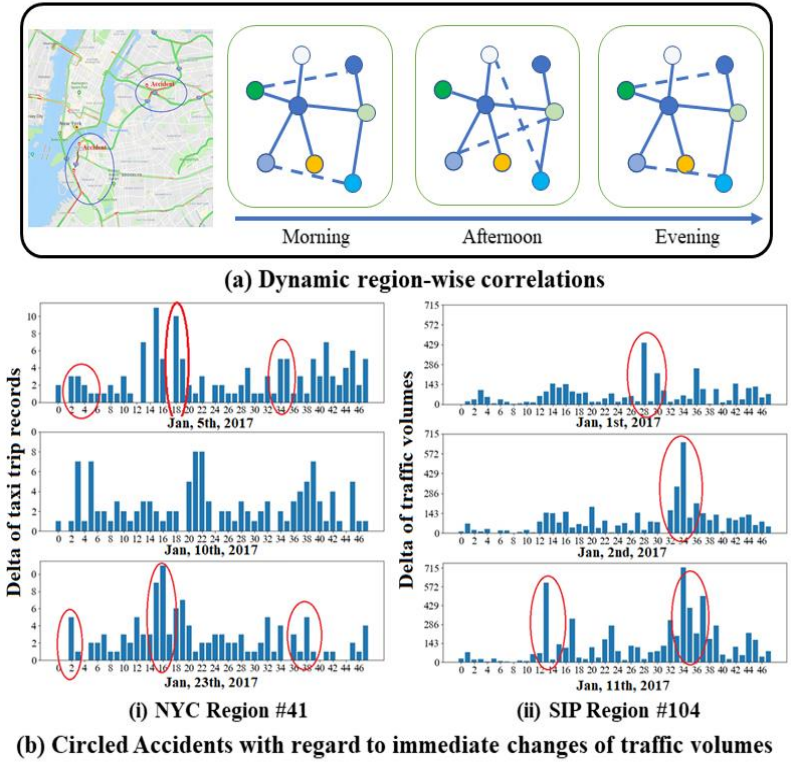
相似的路网结构/
动态车流模式

事故共现
Concurrence

时空建模：DT-GCN

（1）强烈的时变特性：潮汐车流产生的动态区域关联。

（2）交通状态突变与事故关联：速度流量在短期内的突变往往指示了事件发生。



时空建模：DT-GCN

（1）强烈的时变特性：区域之间交通模式存在一定的相似性和关联性，这种关联因潮汐车流等原因产生的会随时间变化的不同关联程度 -> **时变亲和度矩阵**

Overall affinity matrix：

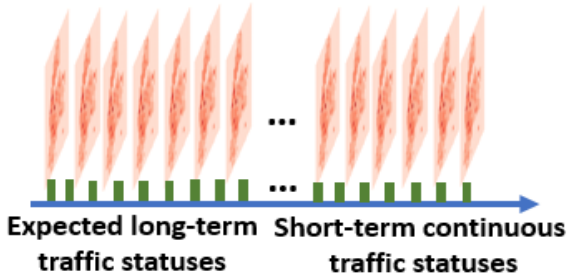
$$\alpha_o^{\Delta t}(i,j) = e^{-JS(s_i^* \| s_j^*)} + \gamma * e^{-JS(C_i^{\Delta t} \| C_j^{\Delta t})} + \beta * tr_{ij}^{\Delta t}$$

静态特征相似性　　动态特征相似性　　区域i到区域j的流量转移

（2）**交通状态突变**：对于同一区域，相邻时间间隔内交通基础元素的数值变化对交通事故的影响（贡献） ->差分特征

$$\overrightarrow{\Theta}^{\Delta t} = \mathcal{D}(\Delta t) - \mathcal{D}(\Delta t - 1)$$

时空建模：DT-GCN



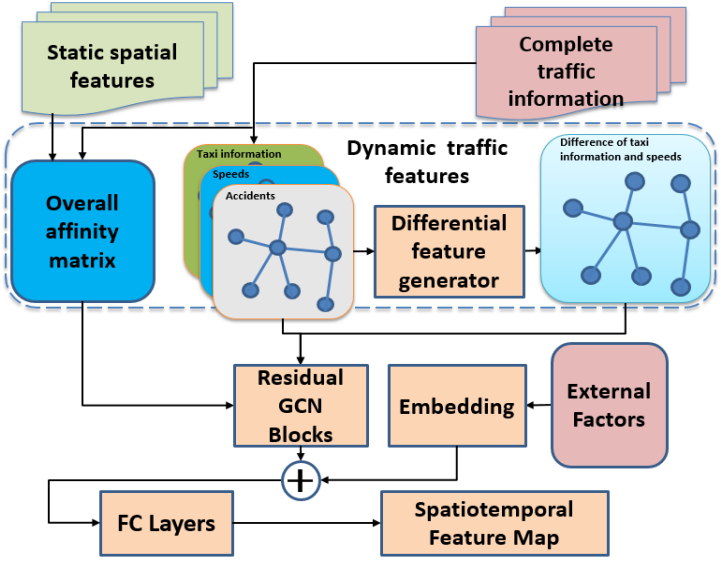
$$\mathcal{B}^{\Delta t} = \mathcal{A}_o^{\Delta t} + I_m$$

$$\mathcal{A}_C^{\Delta t} = \left(\Phi^{\Delta t}\right)^{-\frac{1}{2}} \mathcal{B}^{\Delta t} \left(\Phi^{\Delta t}\right)^{-\frac{1}{2}}$$

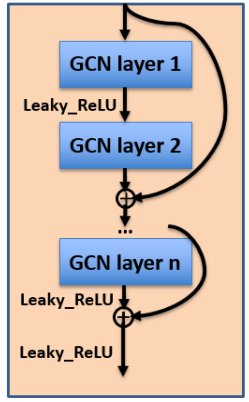
$$\mathcal{U}(\Delta t) = \left\{ \mathcal{F}(\Delta t), \overrightarrow{\Theta}^{\Delta t} \right\}$$

$$\mathcal{H}_{n+1} = \text{Leaky\_ReLU}(\mathcal{A}_C^{\Delta t} \mathcal{H}_n \mathcal{W}_n)$$

$$\text{where } \mathcal{H}_0 = \mathbb{U}_*^{\Delta t}$$



(a) Architecture of DT-GCN

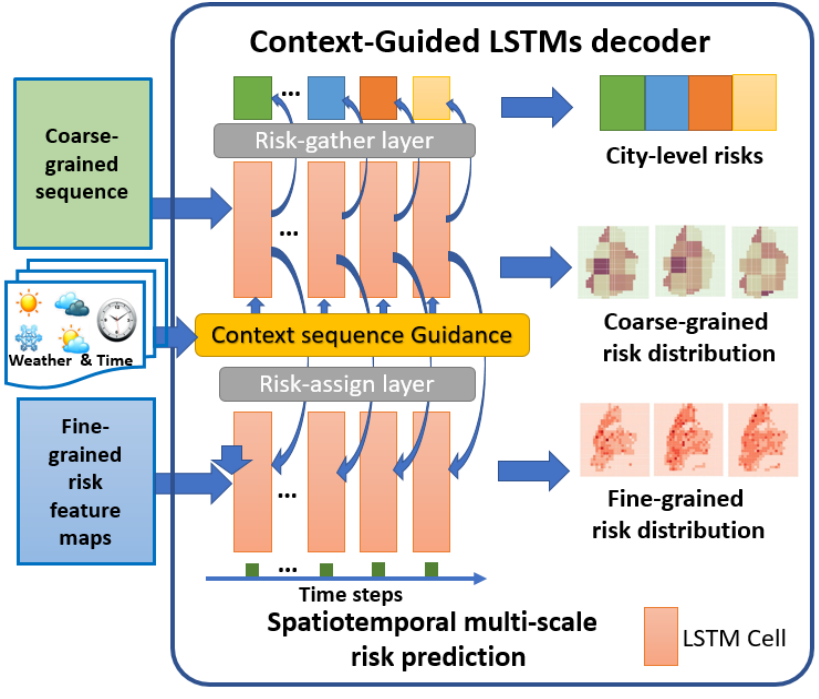


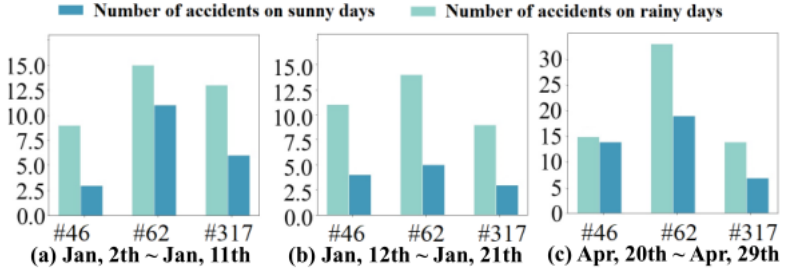
(b) Detailed Residual GCN block

$$\mathcal{M}_F = \{\mathcal{M}_F{}^0, \mathcal{M}_F{}^1, ..., \mathcal{M}_F{}^{h+1}\}$$

RiskSeq: 时空建模 DTGCN & CG-LSTM

时空建模：上下文引导的LSTM（CG-LSTM）





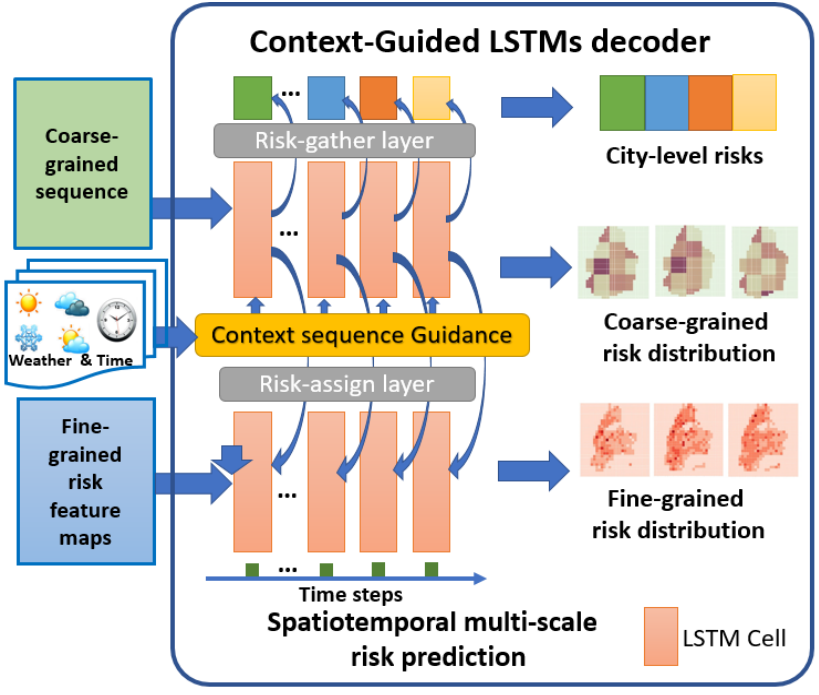
事件稀疏与零膨胀

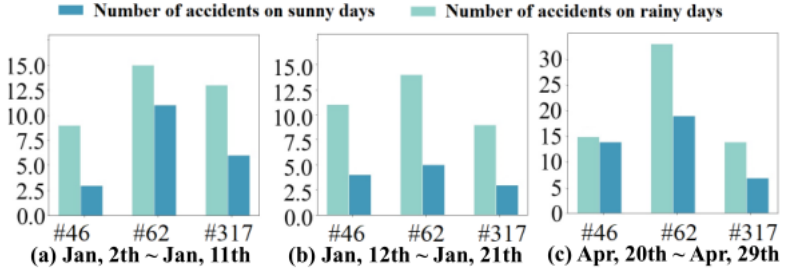
- 空间多粒度的多任务预测
- 粗粒度事故分布视为中间学习信息

空间异质性 多步时序依赖性低

- 引入天气时间上下文逐步引导
- 中间粗粒度风险信息传递至细粒度并融合预测

时空建模：上下文引导的LSTM（CG-LSTM）





$$\mathcal{I}_C^{\Delta t+1} = \mathrm{LSTM_C}(\mathcal{M}_C^{\Delta t+1}, [W_{\mathrm{ext}} * E^{\Delta t+1} + \mathcal{I}_C^{\Delta t}])$$

$$\mathcal{I}_F^{\Delta t+1} = \mathrm{LSTM_F}(\mathcal{M}_F^{\Delta t+1}, [W_{\mathrm{asgn}} * \mathcal{I}_C^{\Delta t} + \mathcal{I}_F^{\Delta t}])$$

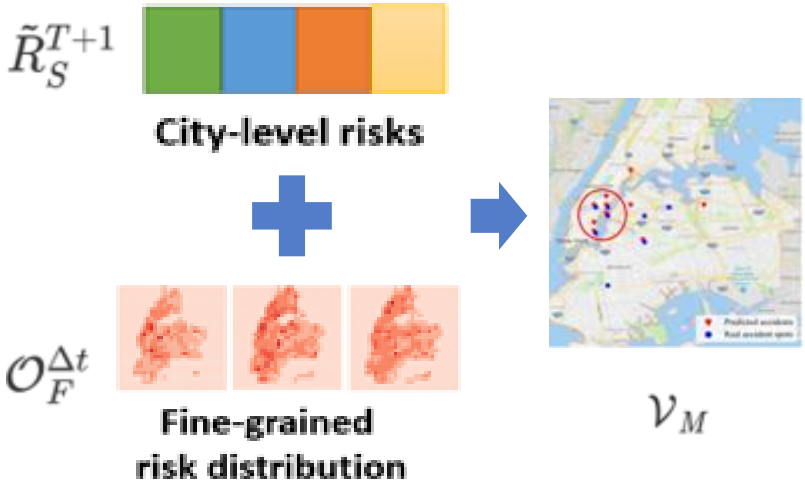
City-level risk $\tilde{R}_S^{\Delta t} = W_{gath} * \mathcal{I}_C^{\Delta t}$

Coarse risk $\mathcal{O}_C^{\Delta t} = \mathrm{ReLu}(W_{CF} * \mathcal{I}_C^{\Delta t} + \mathrm{b}_{CF})$

Fine-grained risk $\mathcal{O}_F^{\Delta t} = \mathrm{Leaky\_ReLu}(W_{RF} * \mathcal{I}_F^{\Delta t} + \mathrm{b}_{RF})$

$$\mathrm{Loss}(\theta) = MSE_F + \lambda_1 * MSE_C + \lambda_2 * MSE_R + \lambda_3 * \mathrm{L2}$$

后处理阶段：自适应事故区域筛选



$$\{< \mathcal{O}_F^{T+1}, \mathcal{O}_C^{T+1}, \tilde{R}_S^{T+1} >, ..., < \mathcal{O}_F^{T+r}, \mathcal{O}_C^{T+r}, \tilde{R}_S^{T+r} >\}$$

每一时间步，将全城事故风险视为事故总数，

自主地选择$K(\Delta t) = int(\widetilde{R_S^{\Delta t}})$选取个数阈值

结合细粒度风险值，选取得到Top-K区域

数据集统计值与评估指标

City	Dataset[5]	Time Span	# of Regions	# of Records
NYC	Accidents	01/01/2017-05/31/2017	354	254k
	Taxi Trips			48,496k
	Speed Values			125k
	Weathers			604
	Demographics	Investigated in 2016		195
	Road Network			102k
SIP	Accidents	01/01/2017-03/31/2017	108	183
	Traffic Flows			1,399k
	Speed Values			311k
	Weathers			180

纽约（NYC）与苏州工业园区（SIP）数据集统计指数

评估指标

回归视角：MSE

分类视角：Acc@M 关注Hit到的准确率

选取风险最高的M个区域和真实发生事故区域进行比较

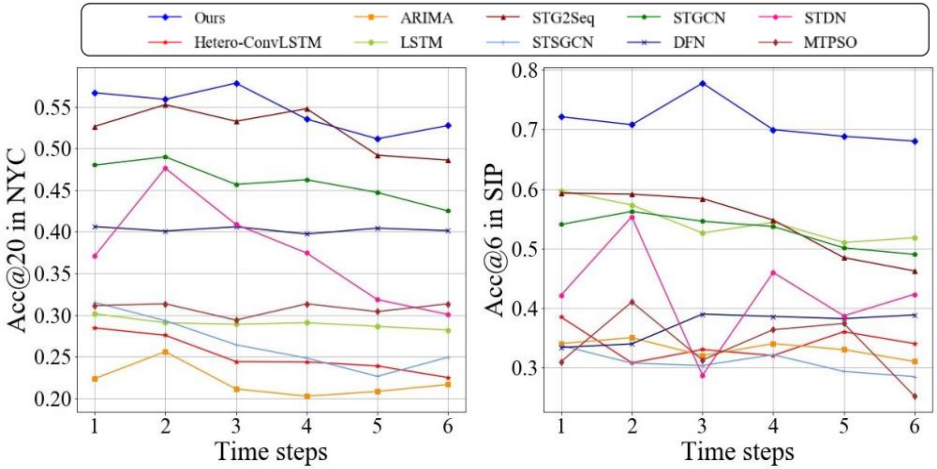
$$Acc@M=\frac{|V_{GT} \cap V_M|}{M}$$

模型横向比较：总体结果与多步预测质量对比

模型横向比较-平均结果

Models	NYC/SIP		
	Acc@20/Acc@6	MSE-F	MSE-C
ARIMA	20.72/30.63	0.0192/ 0.0162	0.0492/0.2215
LSTM	28.98/35.70	0.0179/0.0255	0.0477/0.2694
Hetero-ConvLSTM	28.03/34.84	0.0161/0.0487	0.1015/0.4039
STGCN	50.42/51.27	0.0188 /0.0452	0.0492/0.2885
STG2Seq	52.08/54.30	0.0138/0.0364	0.0693/0.1667
STSGCN	26.46/33.59	0.0183/0.0236	0.1285/0.3473
STDN	37.48/42.18	0.0203 /0.0354	0.0853/0.2142
DFN	40.26/36.98	0.0194 /0.0376	0.0548/0.2278
MTPSO	30.81/33.69	0.0218 /0.0420	0.0393/0.2065
RiskSeq	**56.42/71.27**	**0.0158/0.0401**	**0.0443/0.2702**

模型横向比较-多步预测质量



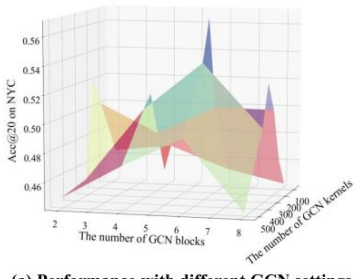
(a) Step-wise performance in NYC (b) Step-wise performance in SIP

模型纵向比较：消融实验与超参数优化

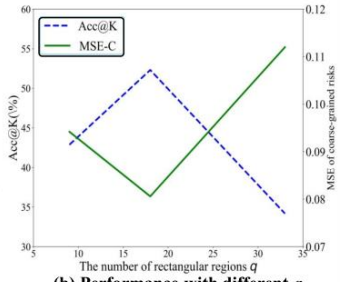
模型纵向比较-消融实验

	NYC/SIP		
Variant	MSE	Acc@20(Acc@6)	Acc@K
RS-PKDE	0.0053/0.0512	18.56/35.48	16.28/29.45
RS-DFM	0.1260/0.0216	43.05/58.94	38.29/46.28
RS-OA	0.0116/0.0127	37.57/67.16	32.47/61.15
RS-DG	0.0118/0.0136	46.45/68.52	39.19/55.27
RS-RC	0.0208/0.0082	41.79/69.45	38.19/56.33
RS-CF	0.0123/0.0355	43.04/67.83	33.21/50.18
RS-CGLSTM	0.0128/0.0060	48.45/67.19	-
Integrated RS	**0.0158/0.0040**	**56.42/71.27**	**47.18/65.26**

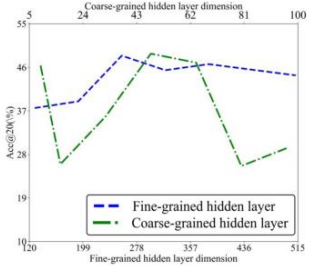
模型纵向比较-超参数网格搜索



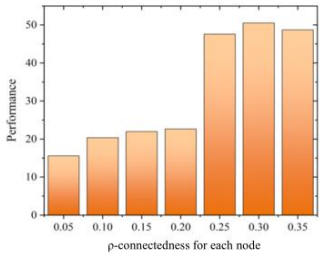
(a) Performance with different GCN settings



(b) Performance with different q



(c) Performance with different LSTM hidden layers



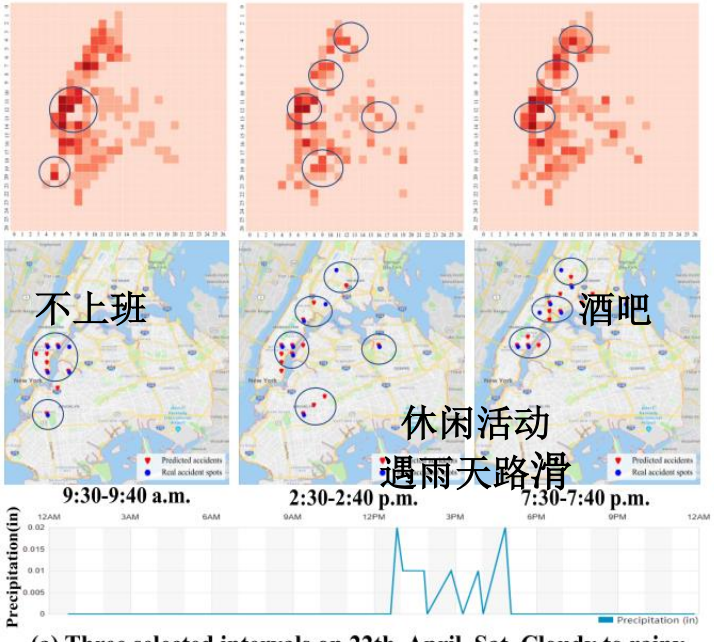
(d) Performance with different ρ-connectedness

案例分析

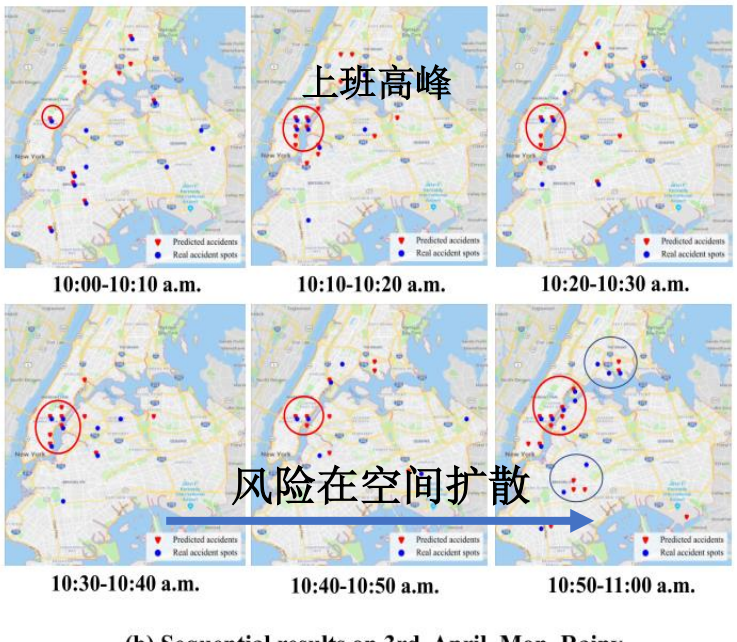
√ 预测事故风险分布大致吻合，筛选结果较好

√ 高风险区域呈现明显的时变特性

√ 预测结果能够跟随上下文变化



(a) Three selected intervals on 22th, April, Sat, Cloudy to rainy



(b) Sequential results on 3rd, April, Mon, Rainy

RiskSeq: 总结与讨论

本文通过**动态聚合邻域内的图信号**以获得更好的风险表示，并采用**逐步的上下文注入和多尺度的时间序列**学习来提升多步事故预测能力。

- **模型可扩展性：**犯罪和流行病预测、推荐系统物品推荐，**偶尔发生并表现出时变的空间依赖和人员流动（人类行为）模式。**DT-GCN与CG-LSTM的多任务可扩展性。

- **稀疏时空数据挖掘：从稀疏性起源的角度**，缓解本质稀疏和伪稀疏问题，将稀疏事件预测转换为可通过DNN解决的**可学习的回归**和**排序任务**。

技术挑战的解决：基于深度学习的时空建模

- 网络结构：图中Edge的时变设计结构、图中signal-wise操作
- 稀疏感知：对问题进行分类解决，以分类的视角设计解决方案
- 困难学习问题：引导机制、多源信息的迁移与半监督学习
- 全局信息与局部信息：车（local）路(global) 协同
- **个体活动随机性与数据稀疏：时空预测不确定性**

拓展延伸：预测准确率瓶颈与时空不确定性

事故时空预测准确率瓶颈

约为55%（击中率=Top-k区域准确预测/所有事故区域）

因：事件“多因一果” 特性，突发性与偶然性

可预测性与不确定性

猜想： 连续时空元素的规律性强？可预测性大？e.g., 速度，流量

离散事件受多种不可控因素影响，可预测性较小？ e.g., 事故，犯罪事件

基于稀疏数据源的预测不确定性大（本质稀疏，伪稀疏，两者并存）？

哪些数据是可预测的，给定的这些多源信息，他的可预测度是多少？

思考：城市时空数据稀疏感知，从稀疏到不确定性

- **人类活动等时空数据蕴藏高度的不确定**
  - 个体活动的不确定：情绪影响、突发事件
  - 环境因素的不确定：自然环境、社会环境
  - 数据采集不确定： 采集过程噪声、数据稀疏
- **数据愈稀疏，可获得信息量愈小，模型难以捕获内在规律，学习过程不确定越大**

- **如何高效利用有限而多源的稀疏时空数据来服务于特定任务？**
- **如何有效捕获时空数据挖掘任务中的时空不确定性及其演变规律？**

心得体会：深度学习应用研究核心要素

1. 数据是什么：输入、输出 – 数据预处理

2. 模型

- 模型结构：数据流动与组合（Concat, LSTM aggregation）
- 损失函数：使模型学习某种信息，进而可以捕获某种物理意义（语义信息）

3. 优化方法： Adam, SGD, AdaGrad等

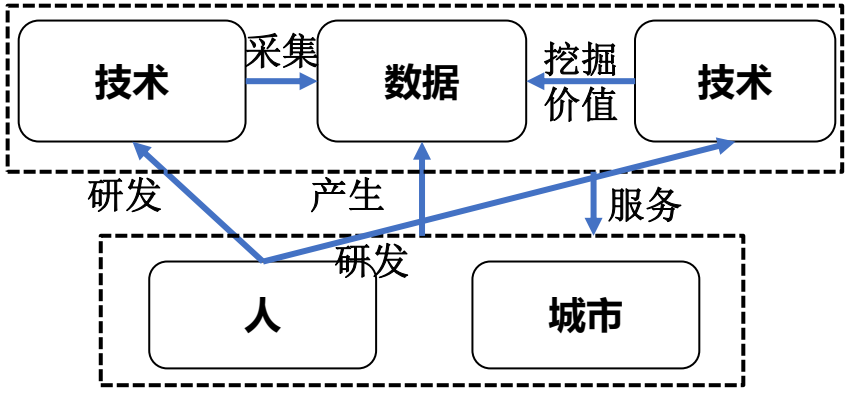
Take away Message

心得体会：论文阅读与写作

- Decoder & Encoder
- 面到点，点到面 即钻进去再钻出来，循环上升
- 不局限自己研究的领域，**广泛阅读->抓共性问题，找相似**

心得体会：技术，人与城市的关系

- 算法、数据服务于人类幸福与城市智慧化
- 技术推动城市治理精细化、科学化、现代化



经济价值

社会价值

事故预测　环境估计
轨迹预测　流量推断
订单匹配　不确定性估计
房价预测　可预测性分析

THANKS!

中科大计算机学院　周正阳

Homepage： http://home.ustc.edu.cn/~zzy0929/Home/

Github: https://github.com/zzyy0929/Codes-for-RiskSeq-TKDE

Email: zzy0929@mail.ustc.edu.cn

Research Interests: Sparse Spatiotemporal data mining

& Semi-supervised Spatiotemporal data learning