A deep learning approach for dynamical systems: a study of train delay prediction in railways

基于深度学习模型的动态系统复杂数据建模方法：
以列车晚点时间预测为例

黄平　Assistant Professor

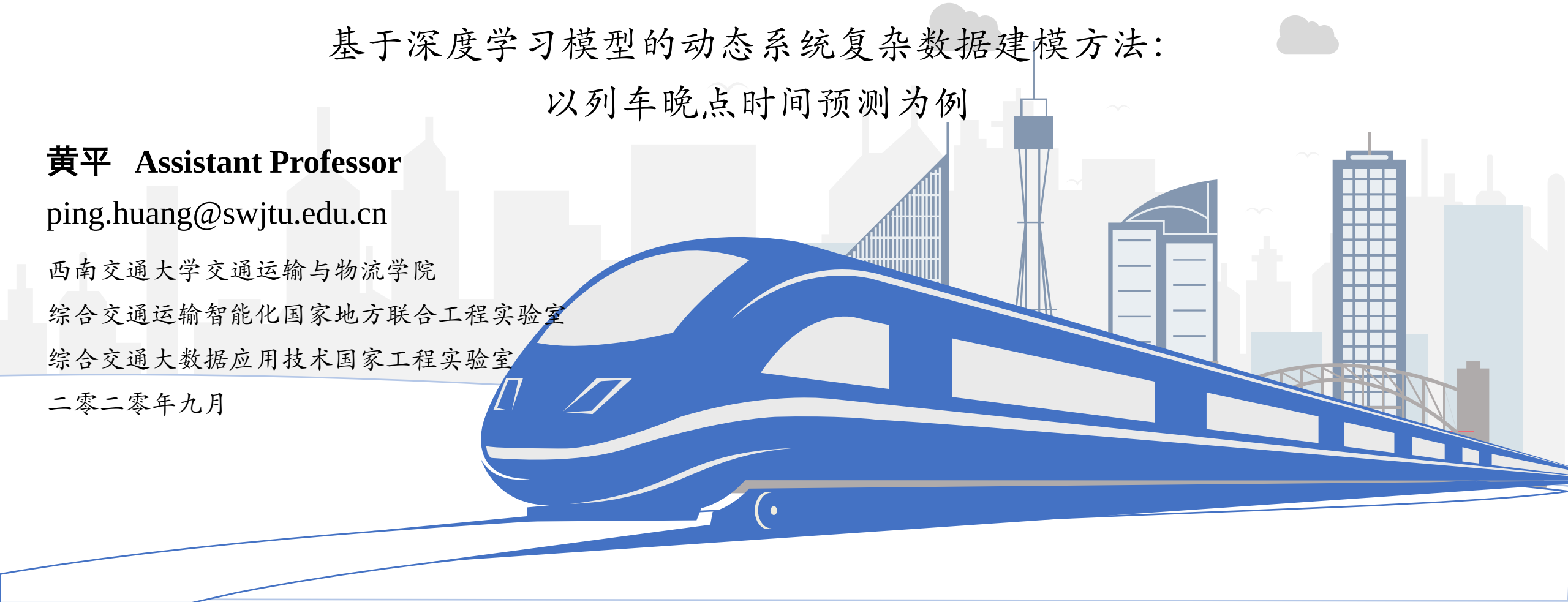
ping.huang@swjtu.edu.cn

西南交通大学交通运输与物流学院

综合交通运输智能化国家地方联合工程实验室

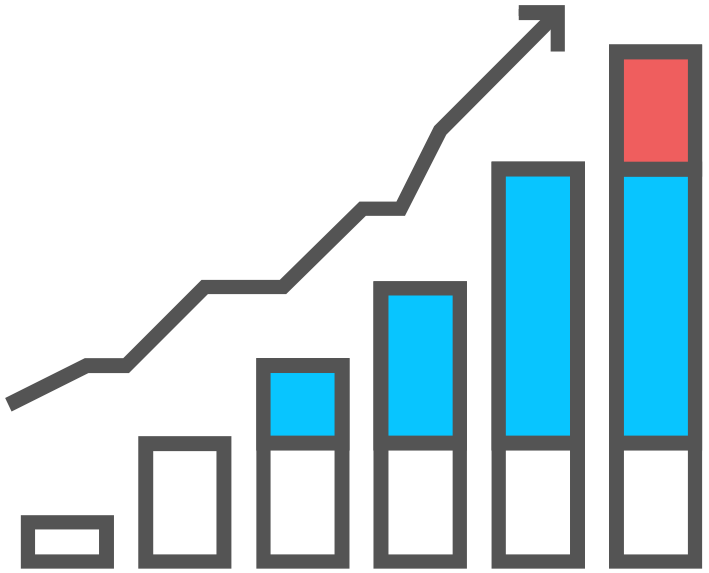
综合交通大数据应用技术国家工程实验室

二零二零年九月



目录
CONTENTS

研究背景与动机



- **传感器**和**通信**等技术使得各系统（自然、社会、经济）内的数据得以大量保存；
- 庞大规模的数据使得我们可以从**大数据的视角**对这些系统进行研究，为我们对这些系统的**管理、控制、预测**等起到了强大的支撑作用；
- 这些系统同时具有如下性质同样使得建模具有一定的难度：

  ✓ 系统影响因素的**复杂性**；

  ✓ 产生数据的**多样性（静态、图像数据、视频数据、时间序列数据等）**；

  ✓ 部分数据具有**趋势、自相关、互相关等特性**；

铁路数据种类

- 列车运行数据
- 客流数据
- 设备状态实时监测数据
- 调度员工作状态实时监测数据
- 基础设施相关数据
- 运行图图片数据
- 视频监控数据
- …

这些数据包含了大量的人肉眼不能发现的信息，借助大数据、人工智能技术将可以挖掘潜在的信息，为实时调度指挥工作提供支撑。

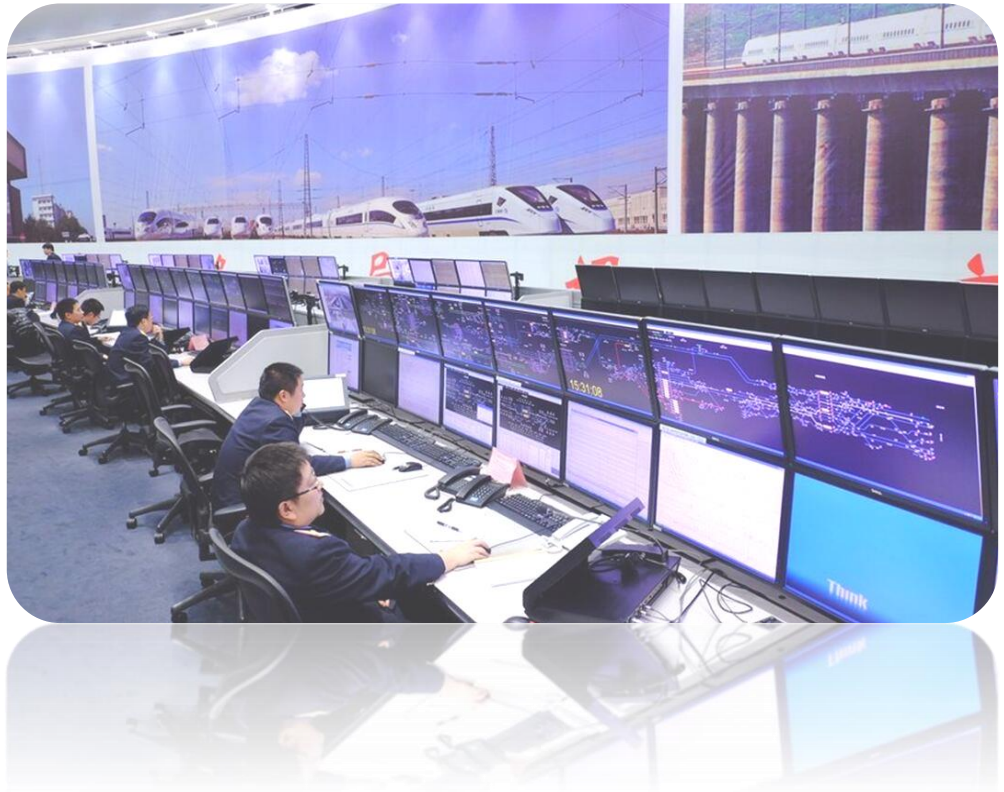
铁路系统内数据特点及其处理方法

■ **动态性、相关性（列车运行数据、天气数据、设备状态数据等）。**

传统主要解决方法：基于统计方法的AR家族模型、MA家族模型、基于机器学习的多层感知器，支持向量机等。

■ **多影响因素属性（人、机器、环境）。**

传统主要解决方法：基于统计分析的多元回归、贝叶斯网络等，以及基于机器学习和深度学习的支持向量机模型、随机森林模型等。



研究背景与动机

- 人工智能和云计算技术的飞速发展为解决这些问题提供了新渠道；

- 不同的**神经网络**已经被提出以解决**特定的问题**：
  - ✓ 循环神经网络（RNN）—序列依赖性；
  - ✓ 卷积神经网络（CNN）—空间依赖性（图像数据）；
  - ✓ 生成对抗网络（GAN）—数据生成；
  - ✓ …

研究背景与动机

铁路系统内，列车晚点时间预测的主要影响因素：

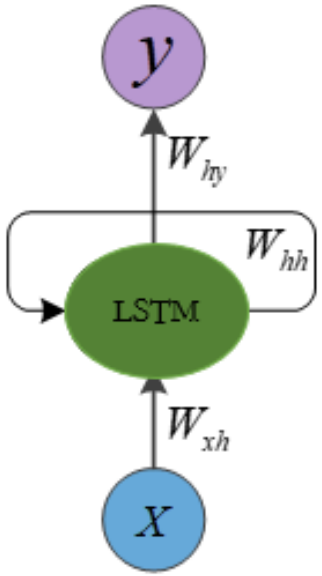
- ◆ 列车历史晚点状态——时空数据
- ◆ 天气、图定运行图相关因素——时间序列数据
- ◆ 设备相关因素（如车站股道、区间长度）——静态数据

本研究结合三种神经网络模型来处理动态系统内产生的多属性复杂数据！

全连接神经网络（Fully connected neural network，FCNN）：

- 在FCNN中，神经元在相邻层之间**完全连接**，信息流从输入层传输到输出层；
- 模型拟合误差被反向从**输出层传播至输入层**（误差反向传播算法）以训练更新各神经元权重与误差；
- FCNN将输入信号作为**静态特征**，其认为输入信号之间没有任何关系，使得FCNN**只能识别自变量和因变量之间的关系**。

长短记忆单元（Long short-term memory，LSTM）是循环神经网络的一种变形：

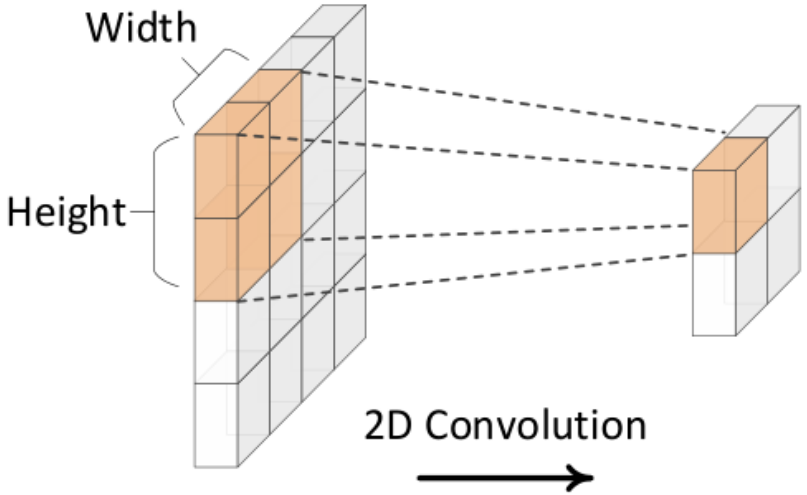


- 常见的序列数据：时间序列、基因序列、语音、句子等。

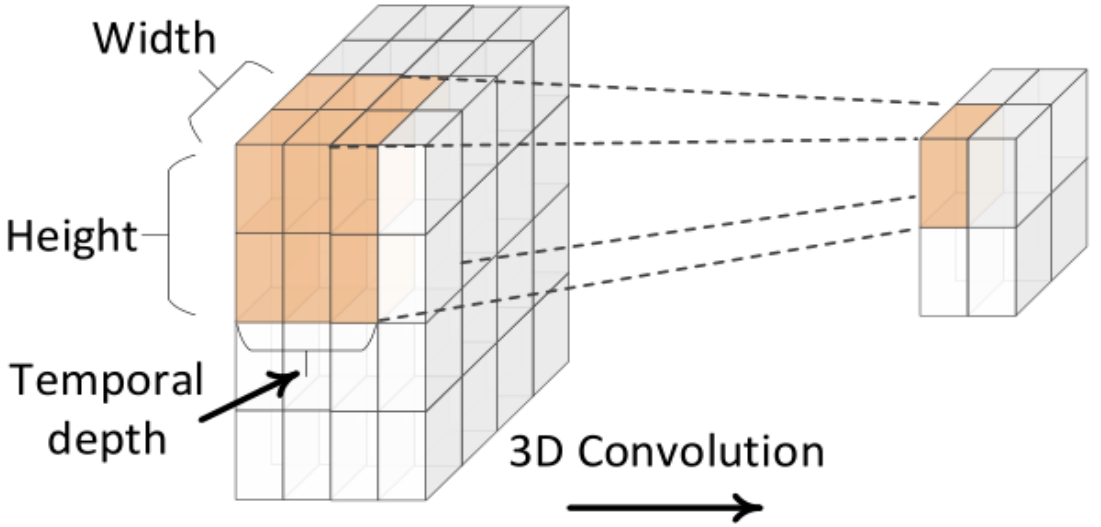
例如一句子序列：　我出生在**中国**，我会说**流利的**____。

- 根据待预测词最近词“流利的”可以大致辨别需要一个语言相关的词语；
- 根据前文的“中国”一词可以推断待预测词语很可能是“中文”；
- LSTM的反馈机制和门控机制使其可以有效地识别待预测词语与前面不同位置词语的相互依赖关系，实现精确的预测。

三维卷积神经网络（3-dimensional convolutional neural networks, CNN）：



2D卷积神经网络（2D CNN）：

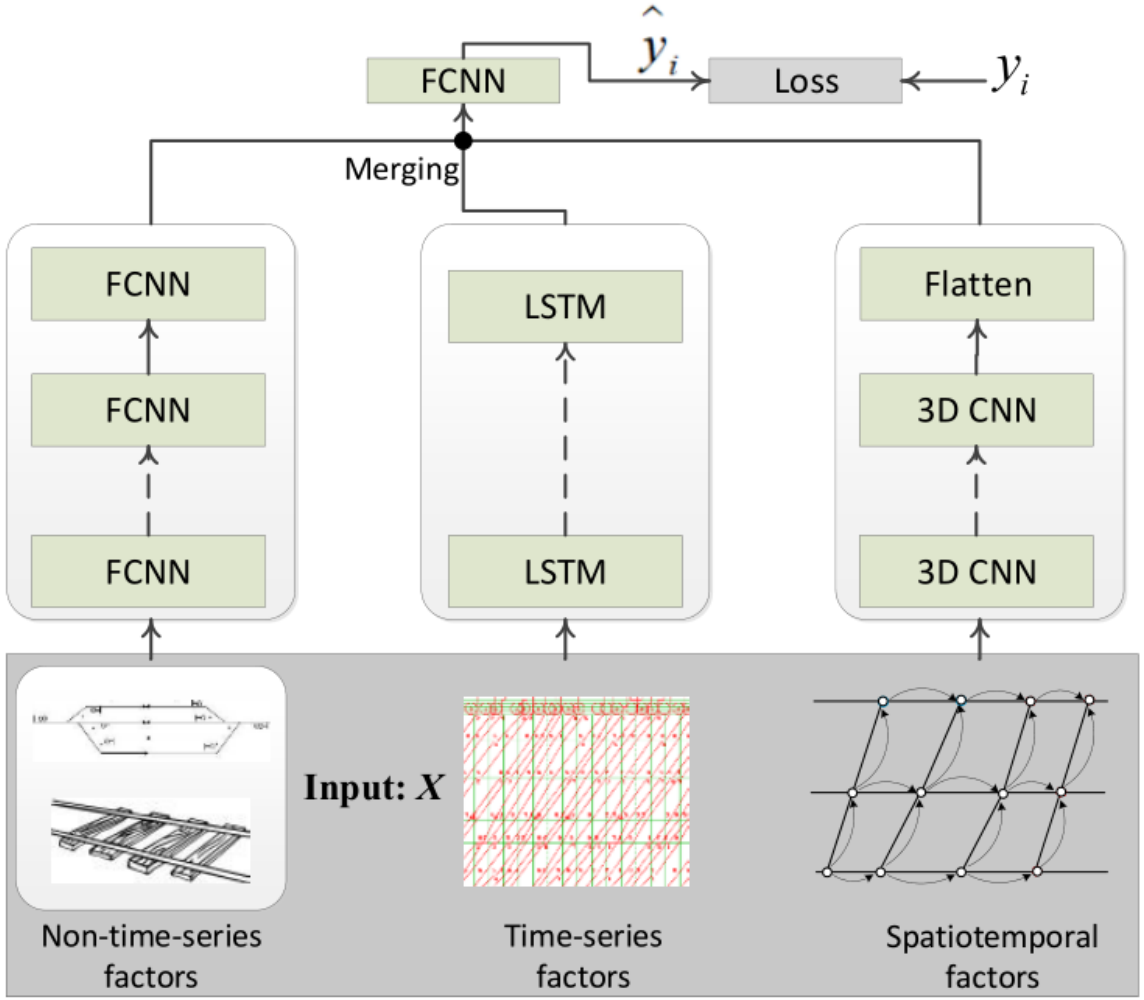


3D卷积神经网络（3D CNN）：

- 3DCNN在CNN的基础上加上了**时间维度**，保留了空间数据上的时间信息，使得其可以同时处理具有时空关系的数据；

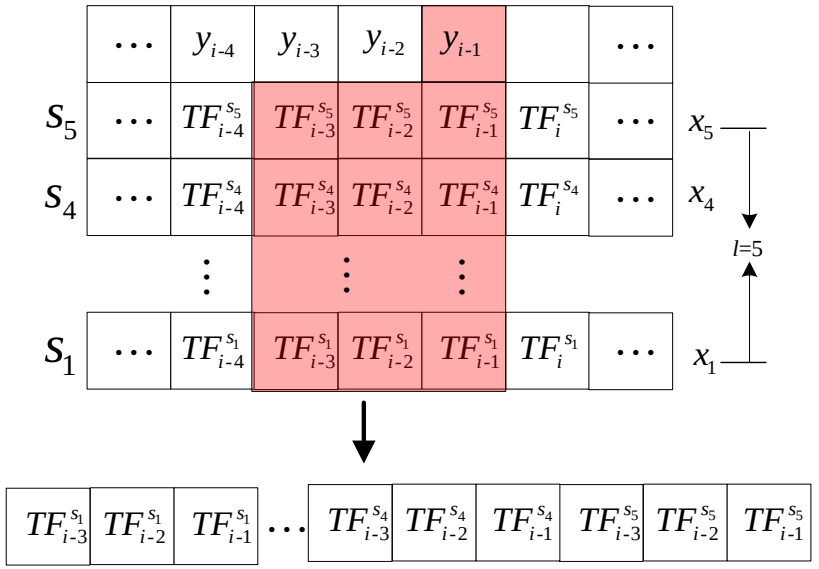
基于FCNN，LSTM，3D CNN的预测模型（CLF-Net）：

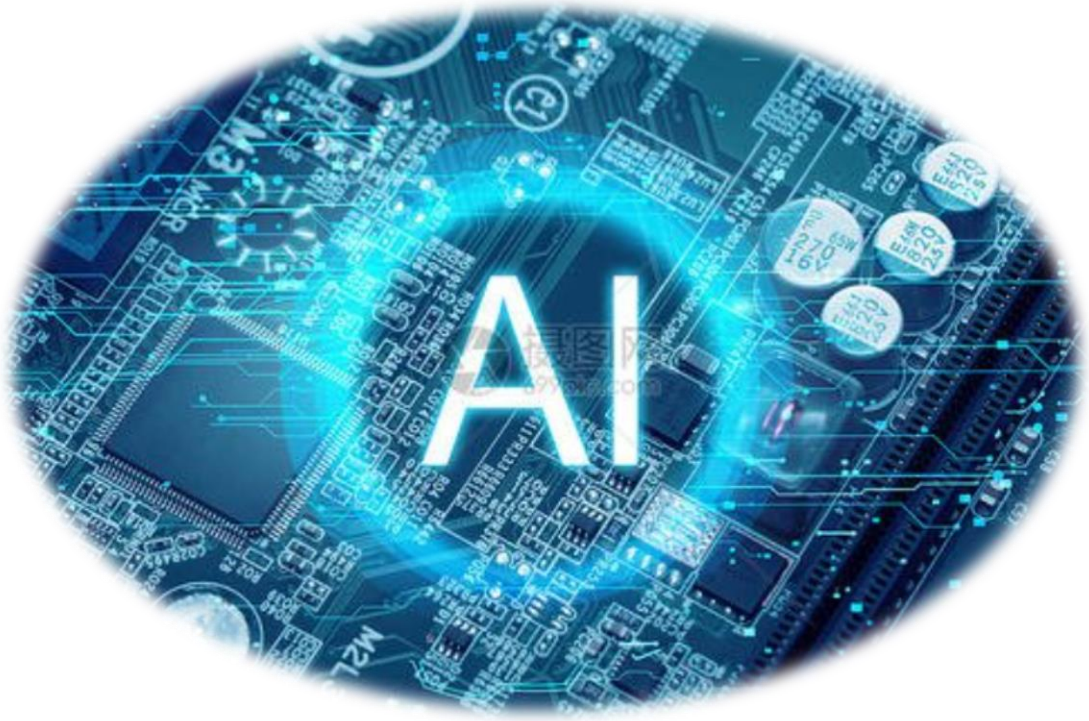
- 在该网络中，每个神经单元执行单独的功能；3D CNN单元被输入**与时空相关性相关**的特征，LSTM单元被输入与**时间序列因子相关**的特征，FCNN单元被馈送**非时间序列特征**。
- CLF-Net中，可以叠加**多个FCNN、LSTM和3D-CNN层**来提高模型的学习能力。



模型介绍

FCNN输入：

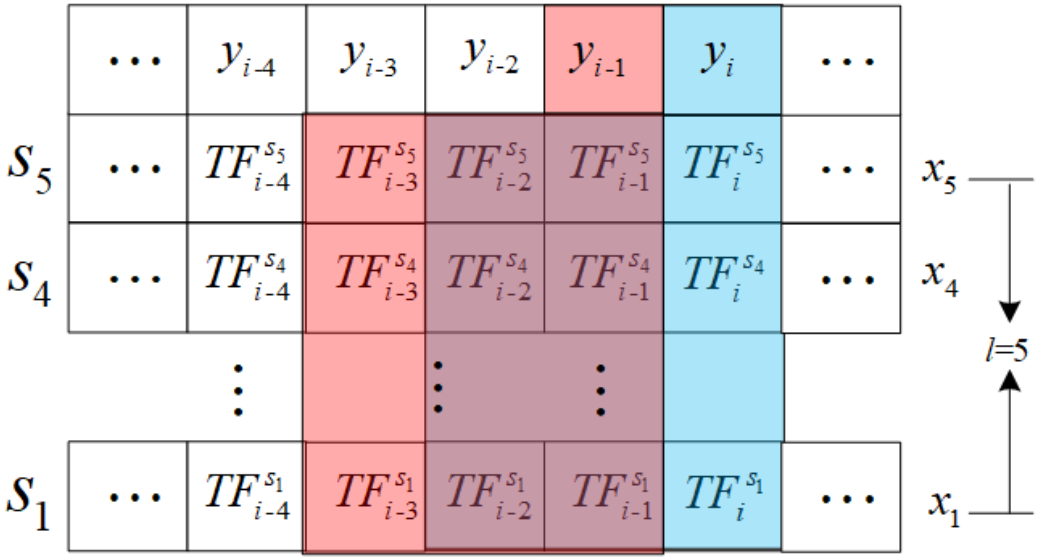




- 将每列车每个车站的特征首位连接形成FCNN输入。

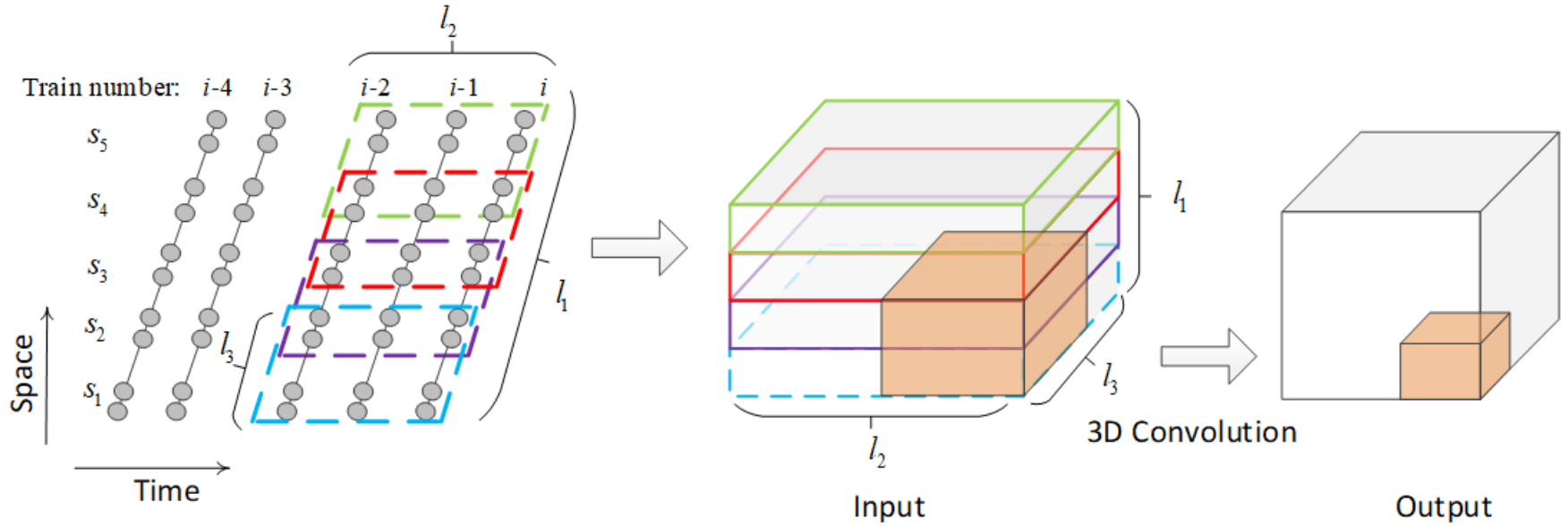
模型介绍

LSTM输入：



- 每个车站代表LSTM的一个时间步。
- LSTM输入为多维时间序列。

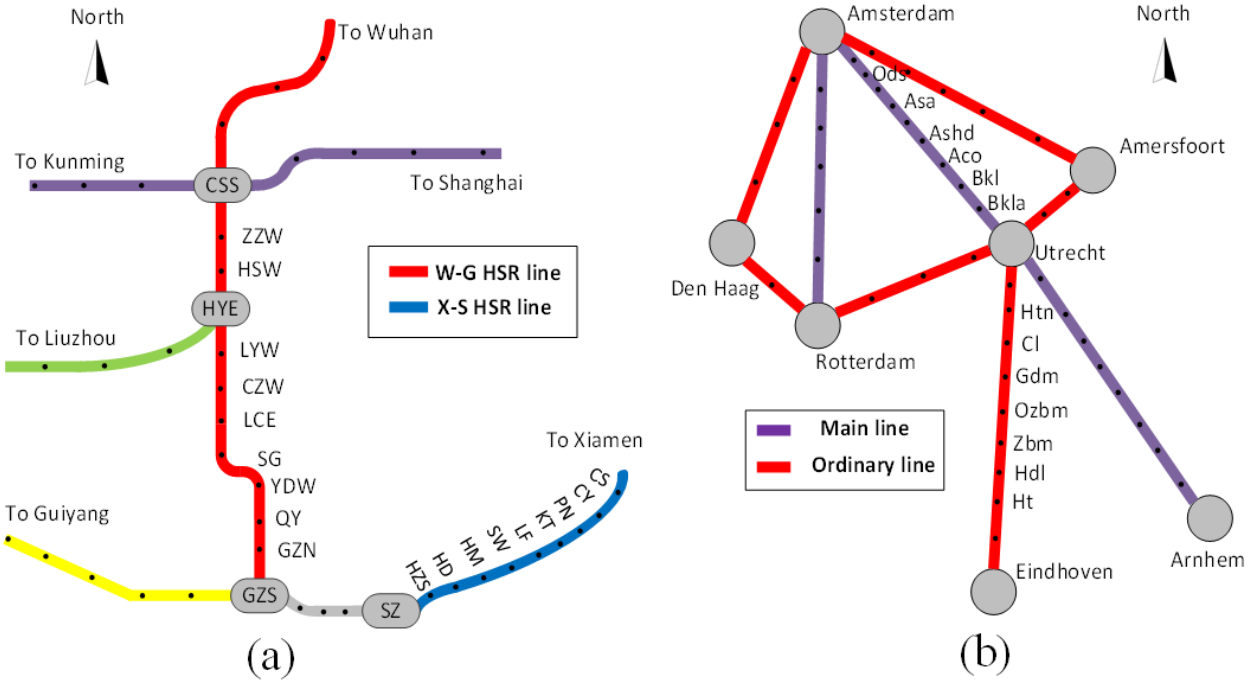
3D CNN输入：



- 2DCNN 在二维空间平面上卷积，3DCNN在三维空间（输入为立方体）上做卷积。

案例分析

- 选用我国**武广、厦深**高速铁路，以及荷兰铁路网**阿姆斯特丹-乌特勒克、乌特勒克-埃因霍芬**铁路运行数据对该模型进行了训练、验证及测试；
- 每条线路选择**4个车站作为预测目标**，提取各变量值，对数据进行**填补缺失值、删除异常值、标准化**等操作；
- 武广（229320个样本），厦深（160650个样本），阿姆斯特丹-乌特勒克（38325个样本）、乌特勒克-埃因霍芬（27825个样本）；
- 随机选取**75%数据作为训练集（模型训练和验证），其余作为验证集及测试集（模型测试）**。



列车晚点时间预测的影响因素：

◆ 时空相关变量（输入到3D CNN）：

1）列车在各站到达晚点时间（min）；
2）列车在各站出发晚点时间（min）；

◆ 时间序列数据（输入到LSTM）：

1）列车在各区间图定运行时间（min）；
2）列车在各区间实际运行时间（min）；
3）列车在各站图定停站时间（min）；
4）列车在各站实际停站时间（min）；
5）列车与前行列车的图定间隔时间（min）；
6）列车与前行列车的实际间隔时间（min）；
7）列车停站次数（次）；

◆ 设备相关因素（输入到FCNN）：

1）各区间长度（km）；
2）各站可用道岔数（整数）；

参数优化：

- 对模型深度（隐藏层数）进行了交叉验证优化；选择使验证数据集误差最小的模型参数，即2层3DCNN，2层LSTM，3层FCNN；

Depth	3D CNN	Loss	LSTM	Loss	FCNN	Loss
1st	3D CNN (32)	0.361	LSTM (64)	0.341	FCNN (32)	0.335
2nd	3D CNN (64)	0.341	LSTM (64)	0.335	FCNN (32)	0.337
3rd	3D CNN (128)	0.341	LSTM (64)	0.338	FCNN (32)	0.328
4th	3D CNN (256)	0.343	LSTM (64)	0.336	FCNN (32)	0.331

参数优化：

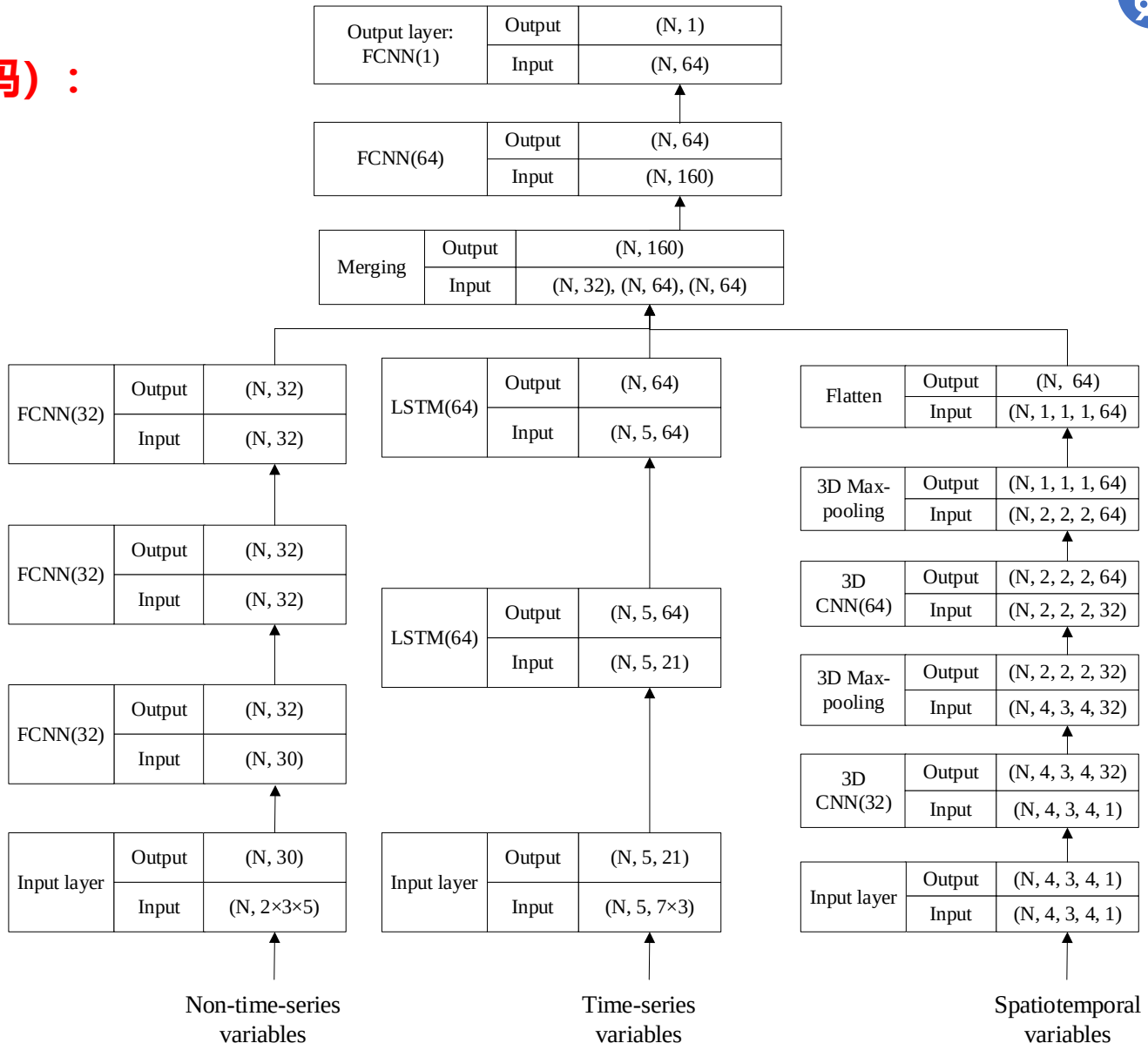
- 对输入维度进行了优化（$l_1 \times l_2 \times l_3$）＝(4×3×4)

l_2	1	2	3	4	5
Validation loss	0.411	0.336	0.328	0.327	0.331

$l_1 \times l_3$	2×9	2×5	3×8	4×4	8×3	5×2	9×2
Validation loss	0.342	0.339	0.334	0.328	0.333	0.337	0.334

- 根据经验选取了卷积核大小(3×3)、激活函数(ReLU)、学习率(学习率衰减法) 等；

模型数据结构图（伪代码）：


Output layer: FCNN(1)
Output (N, 1)
Input (N, 64)
FCNN(64)
Output (N, 64)
Input (N, 160)
Merging
Output (N, 160)
Input (N, 32), (N, 64), (N, 64)
FCNN(32)
Output (N, 32)
Input (N, 32)
FCNN(32)
Output (N, 32)
Input (N, 32)
FCNN(32)
Output (N, 32)
Input (N, 30)
Input layer
Output (N, 30)
Input (N, 2×3×5)
Non-time-series variables
LSTM(64)
Output (N, 64)
Input (N, 5, 64)
LSTM(64)
Output (N, 5, 64)
Input (N, 5, 21)
Input layer
Output (N, 5, 21)
Input (N, 5, 7×3)
Time-series variables
Flatten
Output (N, 64)
Input (N, 1, 1, 1, 64)
3D Max-pooling
Output (N, 1, 1, 1, 64)
Input (N, 2, 2, 2, 64)
3D CNN(64)
Output (N, 2, 2, 2, 64)
Input (N, 2, 2, 2, 32)
3D Max-pooling
Output (N, 2, 2, 2, 32)
Input (N, 4, 3, 4, 32)
3D CNN(32)
Output (N, 4, 3, 4, 32)
Input (N, 4, 3, 4, 1)
Input layer
Output (N, 4, 3, 4, 1)
Input (N, 4, 3, 4, 1)
Spatiotemporal variables

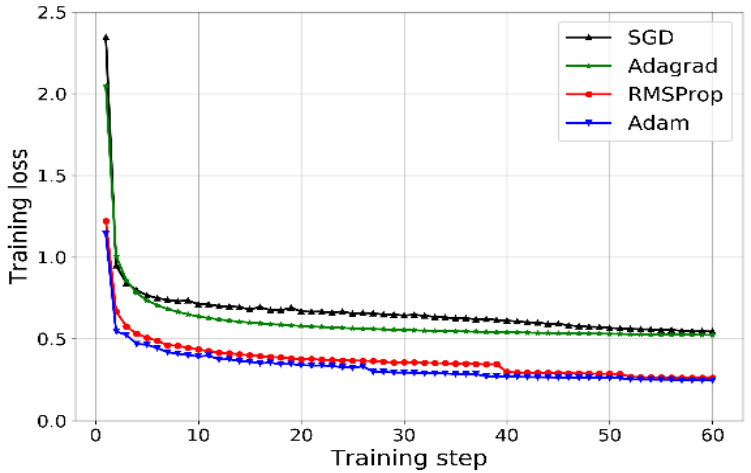
Convergence analysis（模型收敛性分析）：

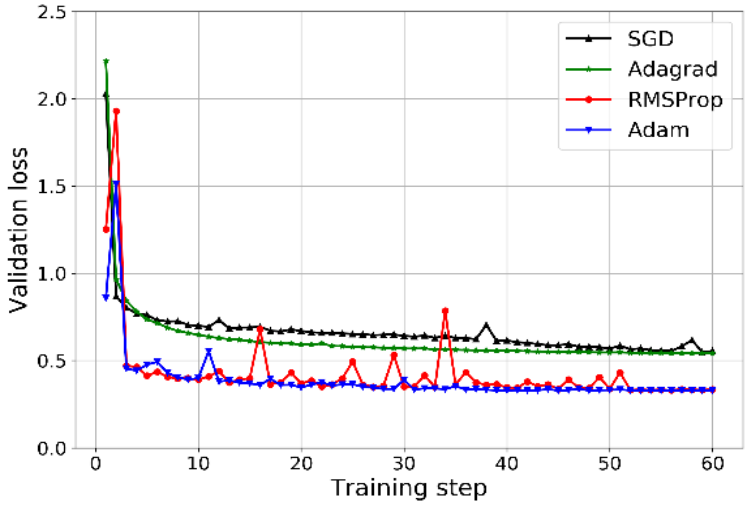
SGD：SGD随机采样部分样本来更新参数；SGDM加入一阶动量；

Adagrad：是一种适用于稀疏梯度的自适应算法；它考虑历史全过程梯度累计效应，可以适应训练过程中参数更新的频率（逐步减小学习率）。

RMSProp：只关心前一段时间的梯度更新频率来更新学习率；

Adam：集成RMSProp和SGDM的优点。





选取了基准模型用于衡量CLF-Net的精度：

- 马尔科夫（MM）
- 多层感知器（MLP）
- 支持向量回归（SVR）
- 随机森林（RF）
- 从CLF-Net中去掉LSTM的模型（3D CNN+FCNN）
- 从CLF-Net中去掉3D CNN的模型（LSTM+FCNN）
- 用CNN-LSTM替换CLF-Net中3D CNN的模型（CNNL-L-F）
- 用Conv-LSTM替换CLF-Net中3D CNN的模型（ConvL-L-F）

选取了RMSE和MAE指标计算各模型预测误差：

$$RMSE=\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i-y_i\right)^2\right]^{\frac{1}{2}}$$

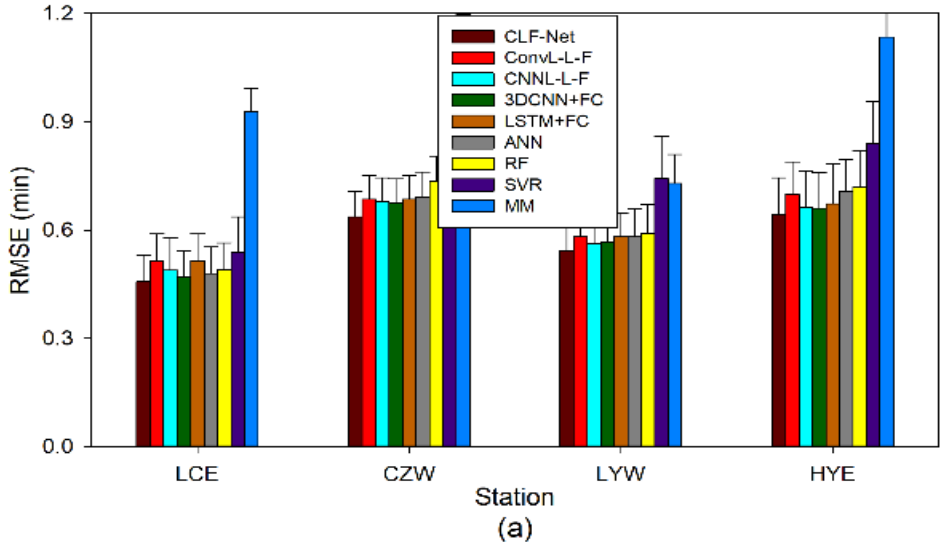
$$MAE=\frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i-y_i\right|$$

案例分析

模型精度评估：

Model	RMSE	MAE	RMSE#	MAE#
MM	0.860	0.644	0.997	0.755
SVR	0.785	0.503	1.071	0.703
RF	0.640	0.443	0.845	0.581
MLP	0.621	0.438	0.788	0.553
LSTM+FC	0.604	0.421	0.786	0.543
3DCNN+FC	0.598	0.418	0.777	0.527
CNNL-L-F	0.603	0.430	0.783	0.542
ConvL-L-F	0.624	0.457	0.782	0.550
CLF-Net	**0.574**	**0.402**	**0.743**	**0.509**

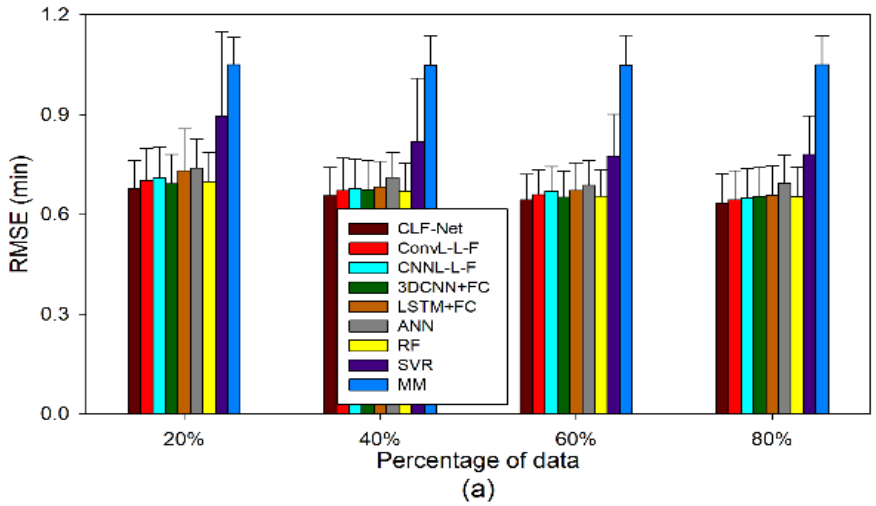
武广线所有车站预测结果

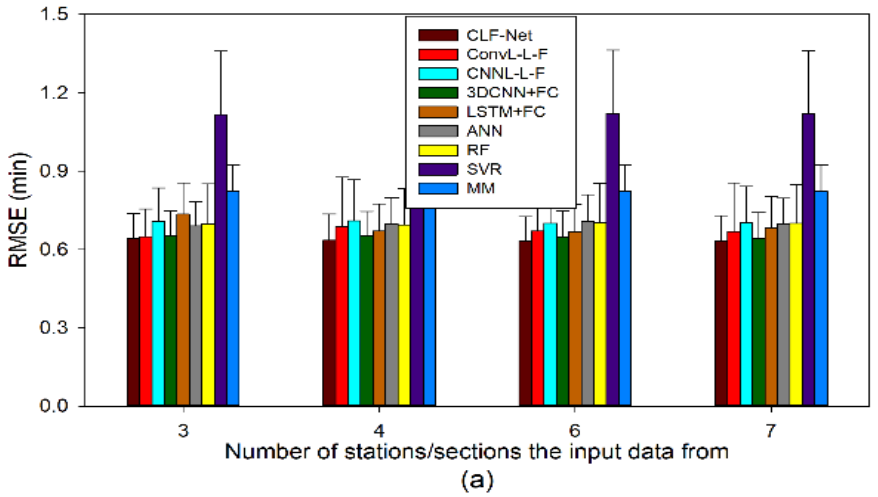


武广线各车站预测结果

模型鲁棒性分析：

- 评估了模型对**不同训练数据量**的适应能力：分别利用武广高速铁路总数据集的20%、40%、60%、80%数据集对模型进行训练和测试；
- 评估了模型对**不同数据维度**的适应能力：以衡阳东（武广线车站）晚点预测为目标，选用不同的**历史数据（输入数据来自过去3/4/6/7个车站和区间）**对模型进行了训练和测试；
- 同时选取前文介绍的8个**模型**作为基准模型，仍然选用RMSE和MAE作为评估指标；





- 该模型**集成了多种神经网络模型**以分别处理不同的数据类型。与传统的机器学习和深度学习模型相比，**该模型具有更好的模式识别和知识发现能力**。

- 模型具有良好的**鲁棒性和扩展性**，可以加入其它类型神经网络模型以适应其它数据类型需求，如道针对路交通中拍摄的图像数据可加入2DCNN模型进行处理。

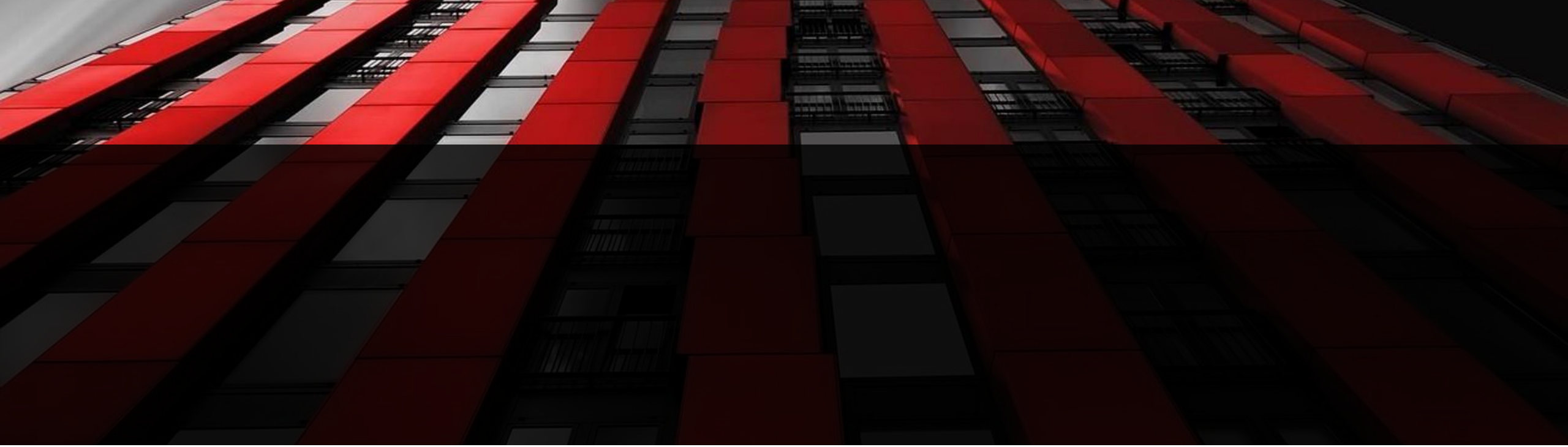
- 最后，模型也可以应用于**其他动态系统**，可以更准确地估计系统内运动对象的状态、位置和轨迹等，从而方便管理者更好地管理和控制这些运动物体。例如，**天文系统中行星的位置和轨迹预测、气象系统中台风的预防以及生态系统中的洪水控制**等等。

[1] **Huang Ping**, Wen Chao, Fu Liping, Peng Qiyuan and Tang Yixiong. A deep learning approach for multi-attribute data: a study of train delay prediction in railway systems [J]. Information Sciences, 2020, 516, 234-253.

[2] **Huang Ping**, Lessan Javad, Wen Chao, Peng Qiyuan, Fu Liping, Li Li, and Xu Xinyue. A Bayesian network model to predict the interruption effects on train operations. Transportation Research Part C: Emerging Technologies, 2020, 114, 338-358.

[3] **HUANG Ping**, WEN Chao, FU Liping, Lessan Javad Jiang Chaozhe, Peng Qiyuan, Xu Xinyue. Modeling train operation as sequences: A study of delay prediction with operation and weather data [J]. Transportation Research Part E: Logistics and Transportation Review, 2020, 141(102022), DOI: 10.1016/j.tre.2020.102022.

[4] **Huang Ping**, Wen Chao, Fu Liping, Peng Qiyuan and Li Zhongcan. A hybrid model to improve the train running time prediction ability during high-speed railway disruptions [J]. Safety Science, 2019, 122. DOI: 10.1016/j.ssci.2019.104510.



Thanks for your attention!