**Deep-ST的一些记录：**

1. **数据部分**

* 每个样本数据为**某一时间点**出租车在网格区域的流动分布。
* 数据分成**原始数据**和**缓存数据**，程序根据设置的周期T进行采样读取位置参数，结合天气，节假日等特征生成所需要的输入数据。

‘data’ : 一系列的时间片

‘Temperature’: 一系列连续值，（the `i^{th}` value is `temperature` at the timeslot `date[i]`）

‘WindSpeed’: 一系列连续值，（the `i^{th}` value is `wind speed` at the timeslot `date[i] `）

‘Weather’: 一个二阶张量（二维数组），One-hot编码（0-17），表示18种气象状态

* **原始数据集**包括：位置，气象，节假日。

位置数据集是一个4阶张量,n可能就是经过时间片划分得到的样本个数；2代指流入、流出这两个指标；32、32即为i,j。把位置空间划分为32\*32的网格进行处理。

* 其中的**外部特征**（除去时间空间以外）**external feature**维度为**28**。其中一周7天+标志是否工作日+标志是否为节假日+~~17维气象~~+2维温度、~~风速~~。（7+1+1+17+2）

这之中的编码为one-hot编码。即除了温度风速归一化，其他是integer(0,1)编码；

* 缓存数据cache中有11对键值对，训练数据和测试数据各5对+1对外部特征的维度。

其中的五对：

Closeness: ,

Period: ,

Trend：,

External: ,

Timeseries:

* 模型的输入数据结构：训练集和测试集结构类似，个数n不一样。模型输入的数据维度为（4,n,m,32,32），4代表C，P，T和E特征；n为样本个数，m对应近中远参数值\*2，32\*32代表空间坐标。
* **数据部分解码h5**

**使用matlab查看数据形状：**

Eg. BJ13\_M32x32\_T30\_InOut.h5

该数据格式下有一个group，即为根目录；group下有两个文件夹（dataset），其中dataset1：

**Dataset 'data'**

**Size: 32x32x2x4888**

**MaxSize: 32x32x2x4888**

**Datatype: H5T\_IEEE\_F64LE (double)**

**ChunkSize: []**

**Filters: none**

**FillValue: 0.000000**

实质上是一个4阶张量；地图划分为32\*32的网格，流入流出流量为2维，4888指的是时间片并和另外一个文件夹（dataset2）date对应：把一天分为48个时间片，一天有48笔数据；date数据shape如下所示：

**Dataset 'date'**

**Size: 4888**

**MaxSize: 4888**

**Datatype: H5T\_STRING**

**String Length: 10**

**Padding: H5T\_STR\_NULLPAD**

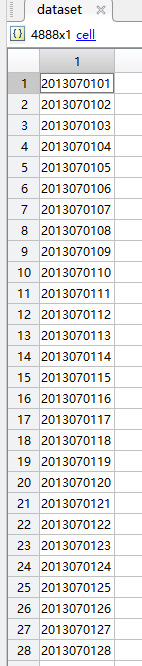
**Character Set: H5T\_CSET\_ASCII**

**Character Type: H5T\_C\_S1**

**ChunkSize: []**

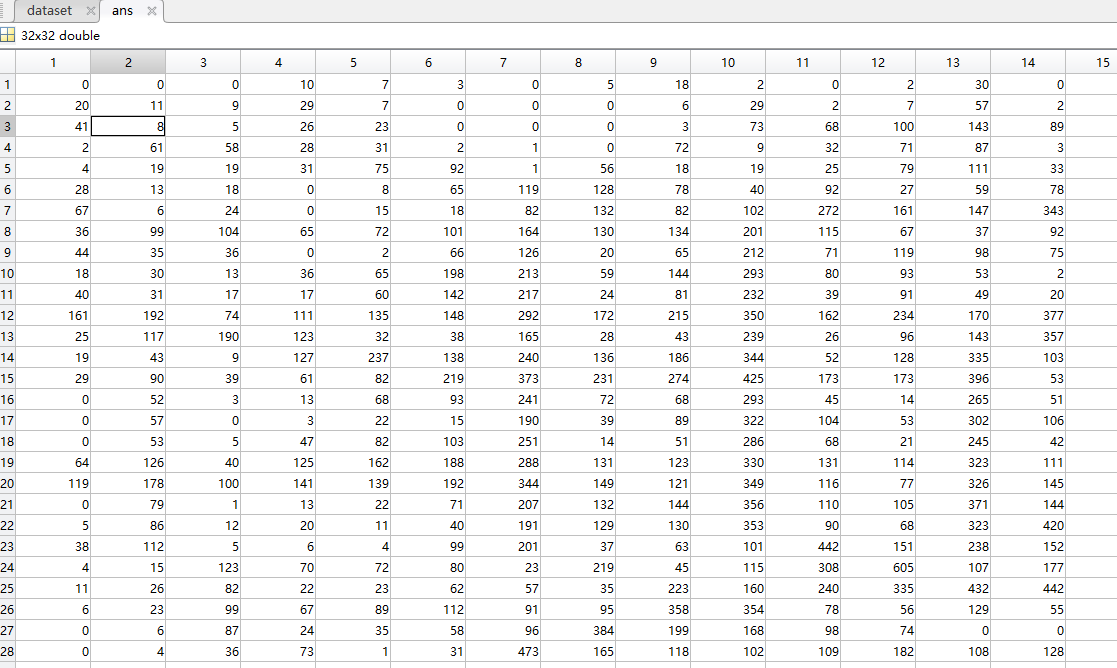
**Filters: none**

**FillValue: ' '**



接着来探索数据：

展示该地区在第一天第一个时间片的流入量（32\*32\*1\*1）：



Matlab操作展示代码：

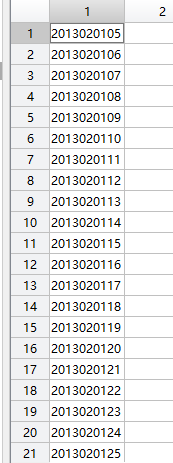
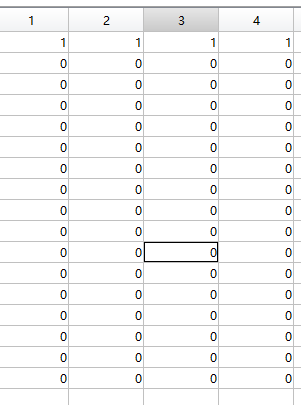
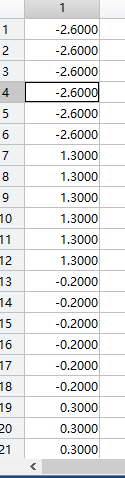
F = 'BJ13\_M32x32\_T30\_InOut.h5';

dataset = h5read(F, '/data');

dataset(:,:,1,1)

**其他张量可视化展示：BJ\_Meteorology**

**date: Weather(one-hot热编码): Temperature:**

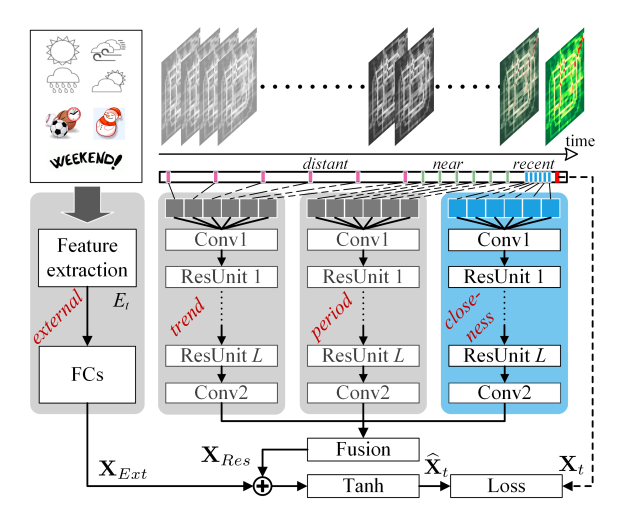
1. **模型部分**

三个角度出发进行建模——时间，空间，额外因素；以Res-net作为模型基础。

* 考虑时间影响：分为closeness影响，period影响，trend影响；
* 考虑空间影响：在CPT三段数据上**分别**采用多层残差神经网络。
* 考虑额外因素影响：利用**全连接神经网络**建模。
* ST-ResNet采用基于卷积的残差网络对城市中任意两个区域之间的远近空间依赖关系进行建模，同时保证模型的预测精度不受神经网络深度结构的影响。

CPT三个模型输出经过加权fuse生成，与相加并通过tanh激活得到。

论文中模型图如下：



CPT三段时间序列分别建立残差网络模型。

模型首先进行一次卷积，然后通过L层残差（每个残差单元有两个卷积层），最后再通过一次卷积。使用adam训练，学习率0.0002，三段时间序列长度分别为3,1,1；每天时刻划分为48份，即半个小时生成一个流动数据图，损失函数为RMSE。卷积核大小3\*3，个数为64。

文章首先解释了为什么不用RNN家族来建模。因为上文说过，这类数据不同于其他时序数据，有周期趋势性，如果要抓住这些特征，需要输入很长的序列，训练会很艰难。

1. **代码部分**

* 模型基于keras；环境已经打包成conda执行文件；
* 进行代码迁移转换的想法记录：

主程序exptTaxiBJ.py中实际引入的脚本文件只有STResNet.py-(stresnet)，TaxiBJ.py，修改代码只需要从这三个文件出发修改即可；首先是stresnet文件，这个是一个做好的时空残差网络的包，具体看到函数stresnet，目前看到需要修改参数：

50： def stresnet(c\_conf=(3, 2, 32, 32), p\_conf=(3, 2, 32, 32), t\_conf=(3, 2, 32, 32), external\_dim=8, nb\_residual\_unit=3):

109：model = stresnet(external\_dim=28, nb\_residual\_unit=12)

回到主函数exptTaxiBJ.py：

29：T = 48 # number of time intervals at a day

39：nb\_flow = 2 # there are two types of flows: inflow and outflow

42：days\_test = 7 \* 4

57+：def build\_model(external\_dim):

171~172（“48”有可能需要修改）：score = model.evaluate(X\_train, Y\_train, batch\_size=Y\_train.shape[0] // 48, verbose=0)

198~199（“48”有可能需要修改）：score = model.evaluate(X\_train, Y\_train, batch\_size=Y\_train.shape[0] // 48, verbose=0)

1. **调参部分**

* **超参数**部分包括：
* nb\_epoch = 500

最开始寻优时的全部参数训练次数，更进一步可以设置模型的fit函数中的patience，一般patience取2~10，代码初始条件为2；

* nb\_epoch\_cont = 100

得到较优模型后继续迭代次数，一般取100。代码初始条件为100；

* **lr = 0.0001**

学习率；一般从0.01开始调整，0.01、0.001、0.0001，代码初始条件为0.0002

* T = 288

采样周期；把24h分为T等份，该条件下为5分钟。

* len\_closeness = 3 # length of closeness dependent sequence
* len\_period = 1 # length of peroid dependent sequence
* len\_trend = 1 # length of trend dependent sequence

时间片长度；文章较为独特的超参数，时间片长度。

* days\_test

预测时间长度；注意与T相乘得整数；

* **nb\_residual\_unit = 2**

残差单元个数；较为重要，设计实验取值{2，4，6，8，10，12}；

* def \_bn\_relu\_conv(nb\_filter, nb\_row, nb\_col, subsample=(1, 1), bn=False):

批标准化；一般而言批标准化结果要好一点。

* def load\_data(T=288, nb\_flow=1, len\_closeness=None, len\_period=None, len\_trend=None,  
   len\_test=None, preprocess\_name='preprocessing.pkl',  
   meta\_data=True, meteorol\_data=True, holiday\_data=True):

外部特征可以选择True或False；一般选True。

* **调参实验待记录部分**
* Learning rate
* **时间片长短。如：len\_trend**
* **时间片间隔。如：TrendInternal**

1. **实验设计部分**

* **自身对照实验部分**

外部特征（True, False）；

不同年份数据集；

预测时间长度（24h,48h,72h; 7天）

残差单元个数；（2，4，6，8，10，12）

* **外部对照实验部分(备选)**

**LSTM**

GRU

RNN

Bi-LSTM

**CNN-LSTM**

**CNN-GRU**

**VAR**

SVM

**ARIMA**

BP

ELMAN

GRNN

CNN-BiLSTM

EMD

ENN

ELM