西南科技大学

高频金融交易中的价格变动预测建模及实现

李若昊 2022年6月1日

研究内容

目的

- 预测某只股票在未来数个周期内的价格变化方向

工具

- 序列模型、卷积神经网络

数据

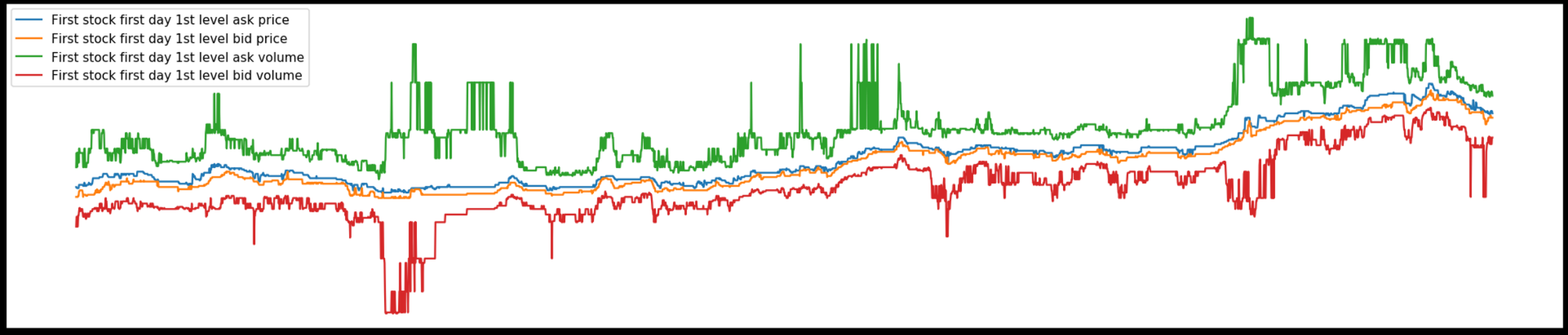
- 公开的高频限价订单簿基准数据集 FI–2010

主要工作

- 建模、调参、训练、可视化

数据来源

- 使用了公开的已标注数据集 FI-2010 中五只股票十档买卖数据合计约40万条快照行情
- 阅读了数据源论文，对原始数据集进行了分析，找出了每只股票在每一天的订单簿中涵盖的范围



模型构建

- 建立了四种类型的模型：

模型种类	各层主要信息
LSTM	100个LSTM单元 + L2正则化 –> Dropout –> 全连接层
CNN(2D)	3个卷积层 –> 最大池化 –> 卷积层 –> 最大池化 –> 全连接层(L2) –> 50% Dropout –> 全连接层
CNN(2D)+LSTM	将CNN中的前一个全连接层替换为了一个含100个LTSM单元的层
CNN(1D)	卷积层 –> 4个空洞卷积层 –> 全连接层(L2) –> 40% Dropout –> 全连接层

模型调优

- 发现对于CNN而言：
  - 它偏好较低的学习速率(≤0.001)和较小的批次大小(≤50)
  - 多于四个卷积层或两个最大池化层均会使预测效果变差
  - 尺寸小一些的滤波器比大一些的效果更好
  - 最优的窗口长度约为每个窗口容纳100个限价订单簿快照
- 对于LSTM来说：
  - 虽然稍大些的批次大小在可接受的范围内降低了预测效果，但使训练效率提升了
  - 尽管对含LSTM单元的中间层用了L2正则化和Dropout，过拟合的现象依然存在；如果想不过拟合得太厉害，只能用单个LSTM层
  - 最优的LSTM层大小也许小于实验中的100个LSTM单元

预测结果

- 各个模型的预测结果如下：

模型种类	Loss	准确率	F1分数	Kappa 系数
LSTM	0.70/0.89	0.72/0.63	0.67/0.63	0.50/0.45
CNN(2D)	0.94/0.98	0.50/0.50	0.38/0.41	0.22/0.25
CNN(2D)+LSTM	0.74/0.83	0.68/0.63	0.68/0.64	0.49/0.45
CNN(1D)	0.41/0.37	0.86/0.88	0.88/0.88	0.82/0.82

可以看到：

- 任何一个模型的效果均优于数据集源论文中给出的基线准确率（F1分数=46%），达到了任务书指定的目标
- 空洞卷积模型的效果远优于其他模型
- 向CNN模型中增添一个LSTM层后提升了它的性能
- CNN+LSTM模型比LSTM模型更不容易过拟合，估计在更长时间的训练之后可以比LSTM模型有更好的结果

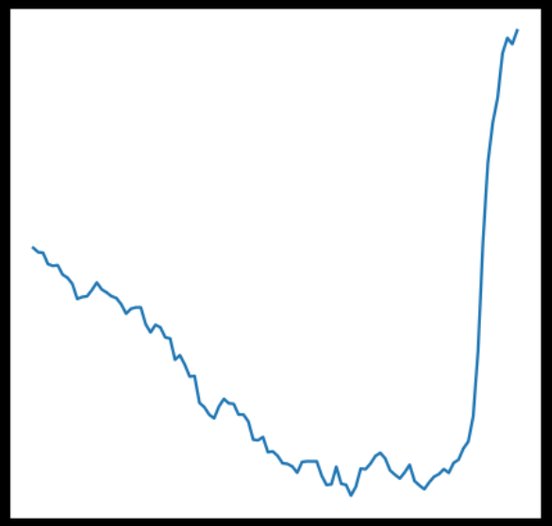
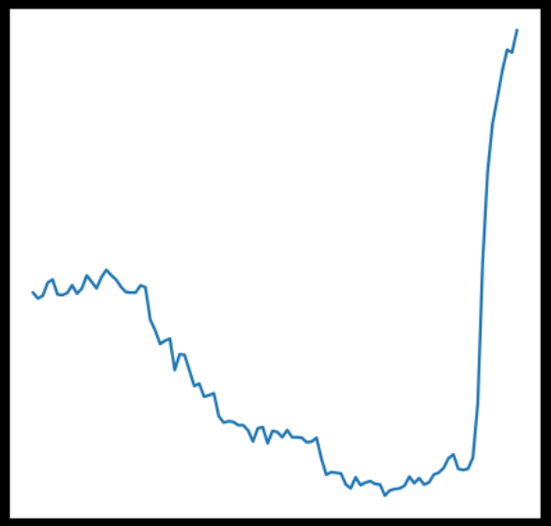
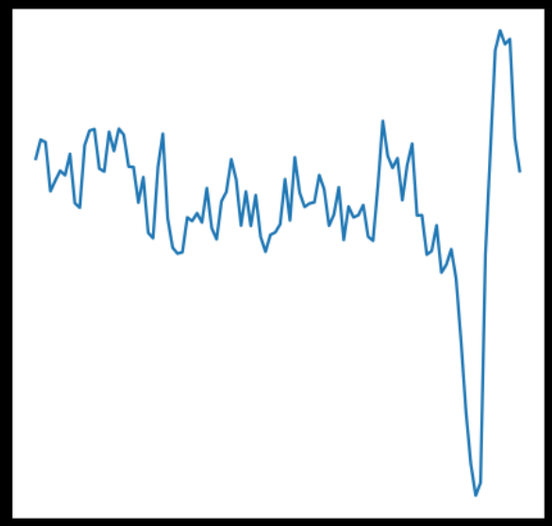
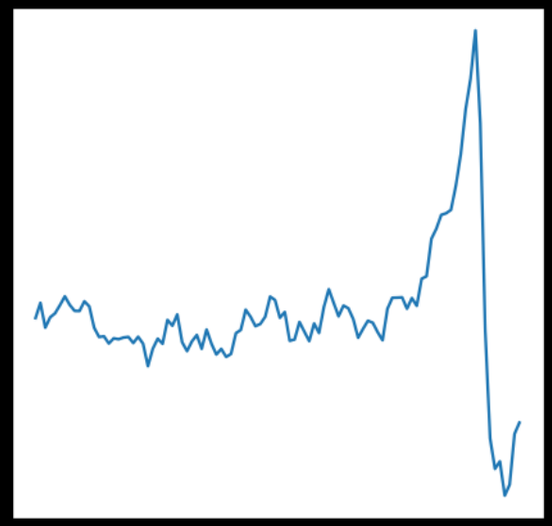
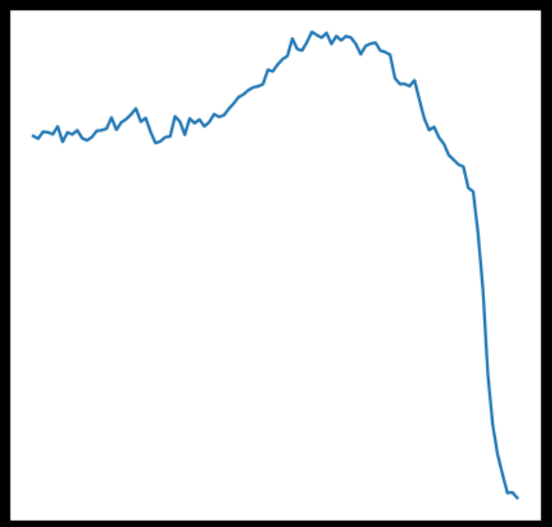
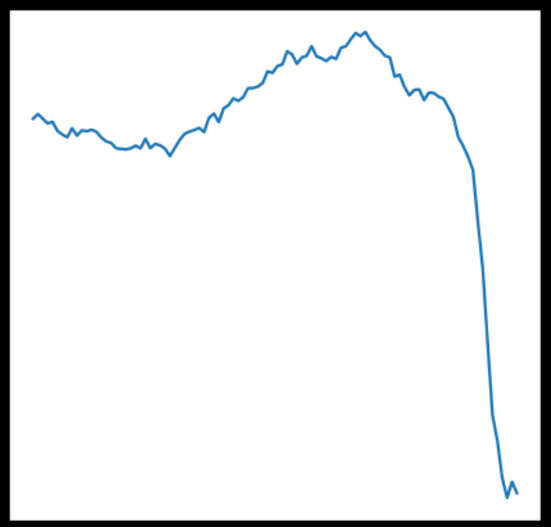
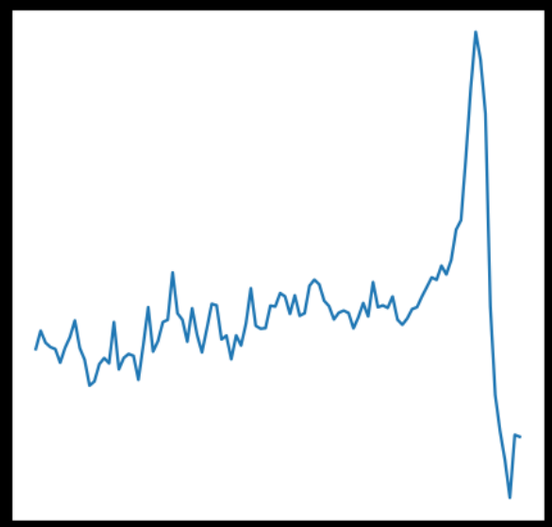
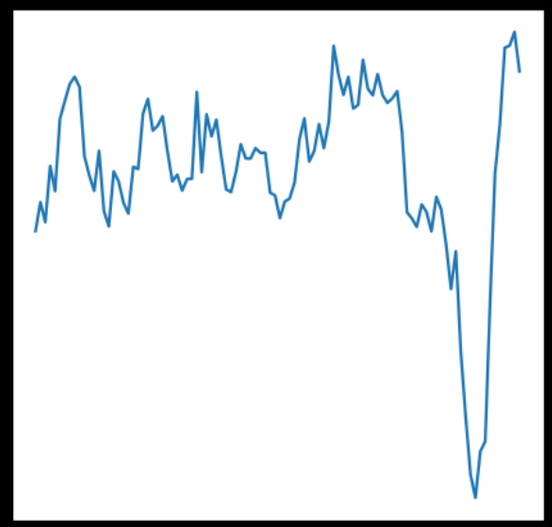
迁移学习

- 比较了5–>1模型和1–>1模型的效果：

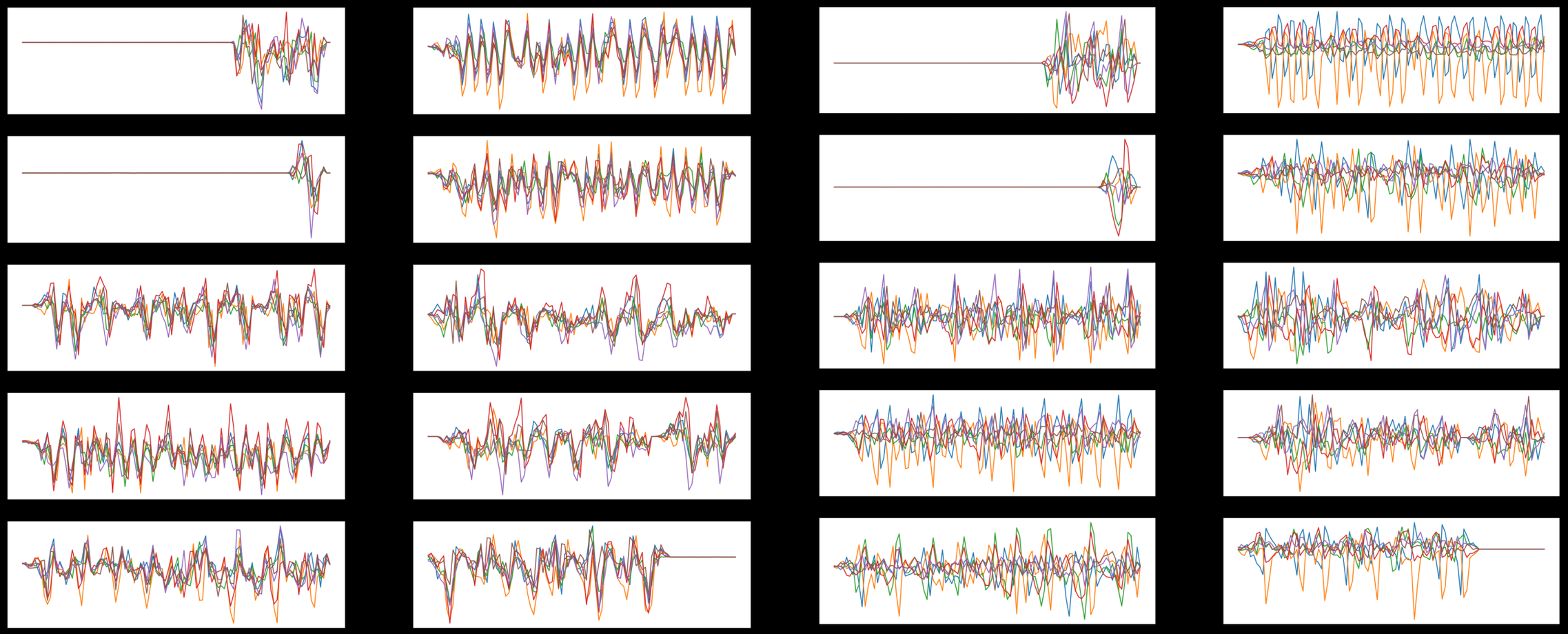
5–>1		1–>1	
F1分数	Kappa 系数	F1分数	Kappa 系数
0.66/0.63	0.47/0.41	0.78/0.63	0.66/0.41

可以发现1–>1模型在测试时指标上与5–>1模型差不多

找出每一批次中得分最高的数据

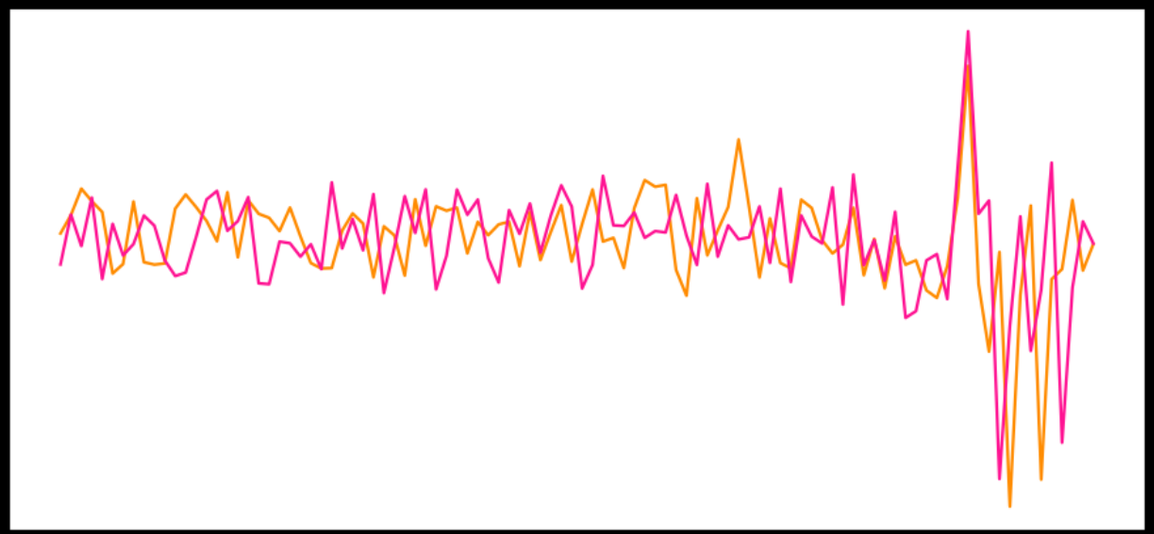









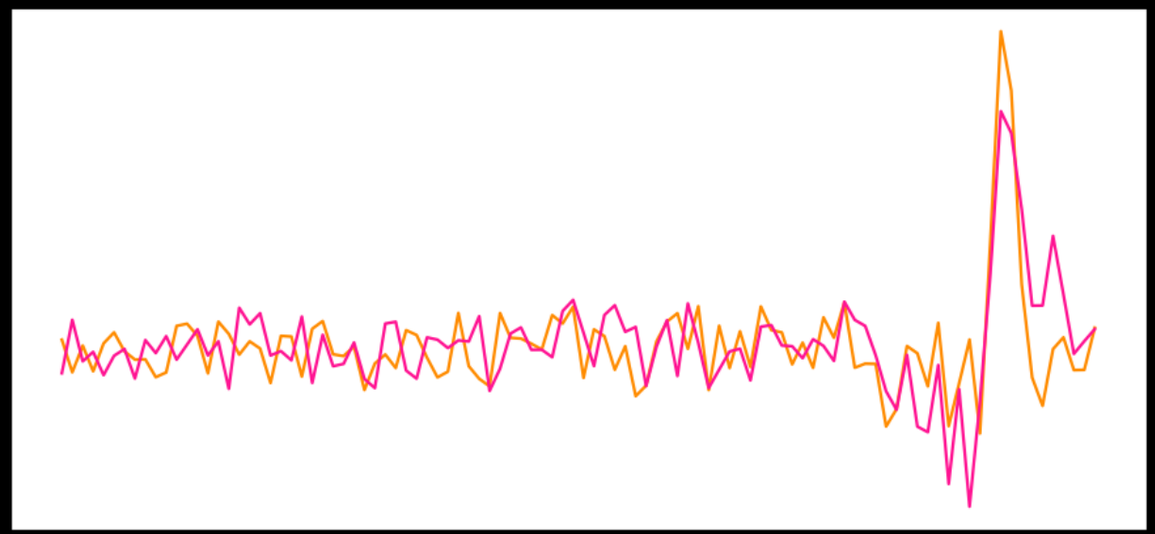
使用梯度上升法可视化了CNN中某个靠后的激活层的滤波器学到了什么



- 看起来在输入中含有价格的某种变化模式时，这一层会积极地作出反应

使用同样的方法可视化了CNN最后的全连接层学到了什么





- 有趣的是买卖价格数据窗口的尾部出现了较高的方向相反的峰，我推测这个现象对应于买卖价格反弹这种微观结构现象

限制

- 空洞卷积模型如此高的准确率在实验时无法在未来5个周期的预测范围内复现
- 目前可视化时间序列的数据方法不如T–SNE等方法好