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6日

研究问题及意义

预测某一只股票在未来数个周期内中间价格的变化方向

且预测准确率要高于指定值

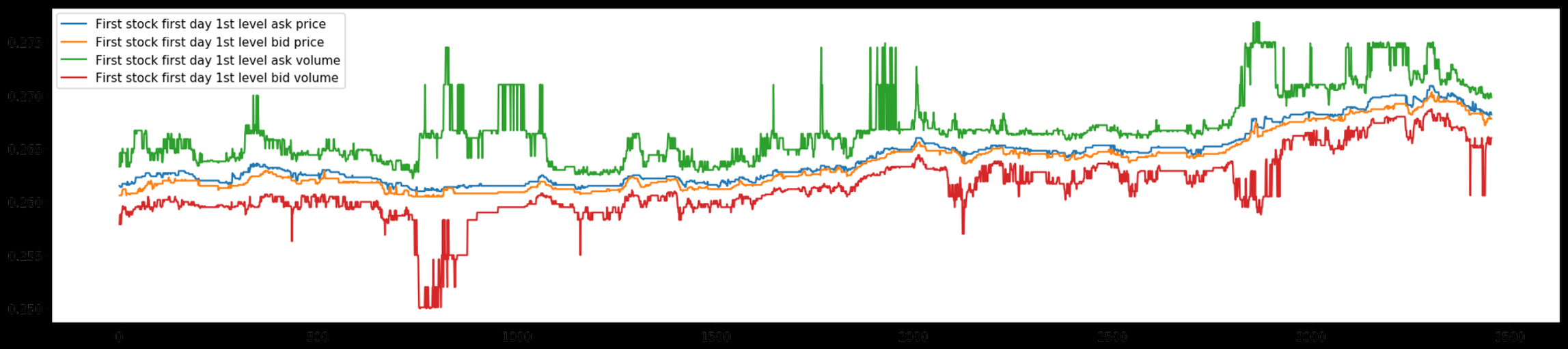
意义：

- 为希望规避逆向选择风险的交易者提供参考
- 为试图甄别非法交易活动的监管机构提供参考

数据源

公开的高频限价订单簿基准数据集 FI-2010（已标注）

- 大规模：约每0.5秒一张快照，合计约40万条委托记录
- 高质量：十档买卖数据



研究方法

序列模型、卷积神经网络

- 监督式学习，对中价变化方向进行分类训练
- 为什么?
  - 数据量大、精度要求高 -> 传统机器学习方法有性能瓶颈

主要工作（一）

- 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

- 自主选择了激活函数、优化算法的种类

主要工作（二）

- 并参考指导老师的经验，大量调优（论文表3–2及附录1）发现：
  - 对于CNN而言：
    - 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

主要工作（三）

- 迁移学习
- 对模型预测原理进行探索

预测效果

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

注：数据格式为验证集结果/测试集结果

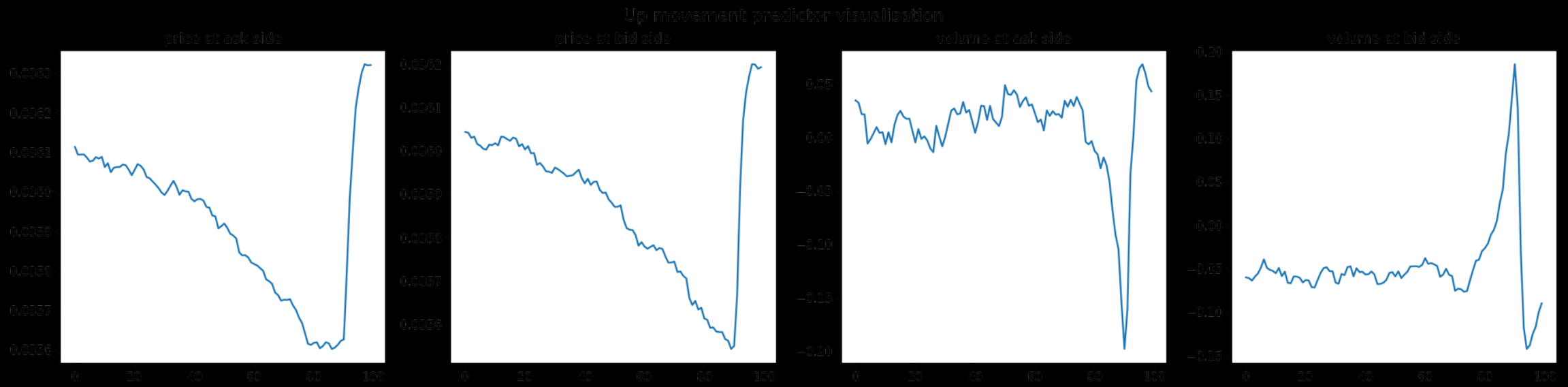
- 任何一个模型的效果均优于数据集源论文中给出的基线准确率（准确率=0.48，F1=0.41），达到了任务书指定的目标
- 空洞卷积模型的效果远优于其他模型
- 向CNN模型中融入一个LSTM层后提升了性能
- CNN+LSTM模型比LSTM模型更不容易过拟合，估计在更长时间的训练之后可以比LSTM模型有更好的结果

迁移学习

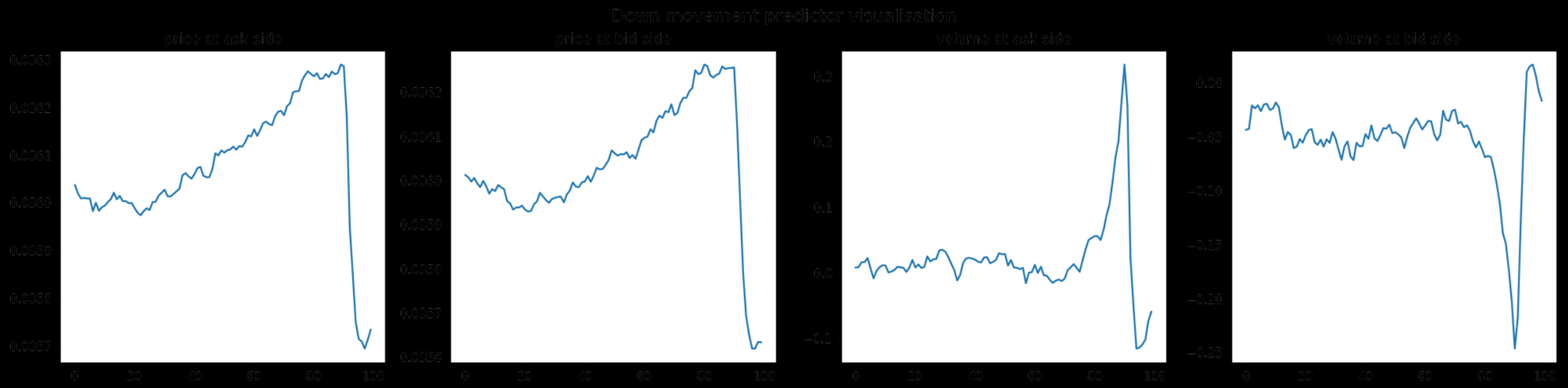
5–>1		1–>1	
F1分数	Kappa 系数	F1分数	Kappa 系数
0.66/0.63	0.47/0.41	0.78/0.63	0.66/0.41

1–>1模型测试时在指标表现上与5–>1模型差不多

每一批次中得分最高的数据

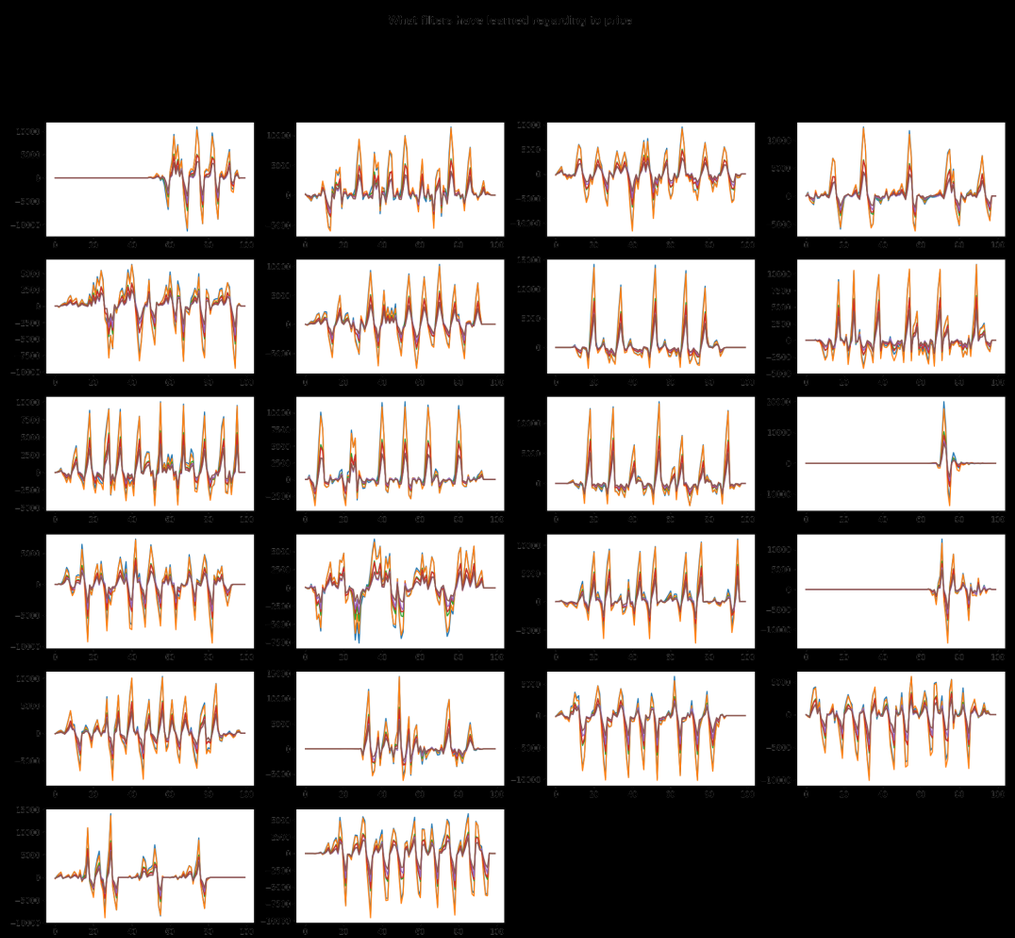


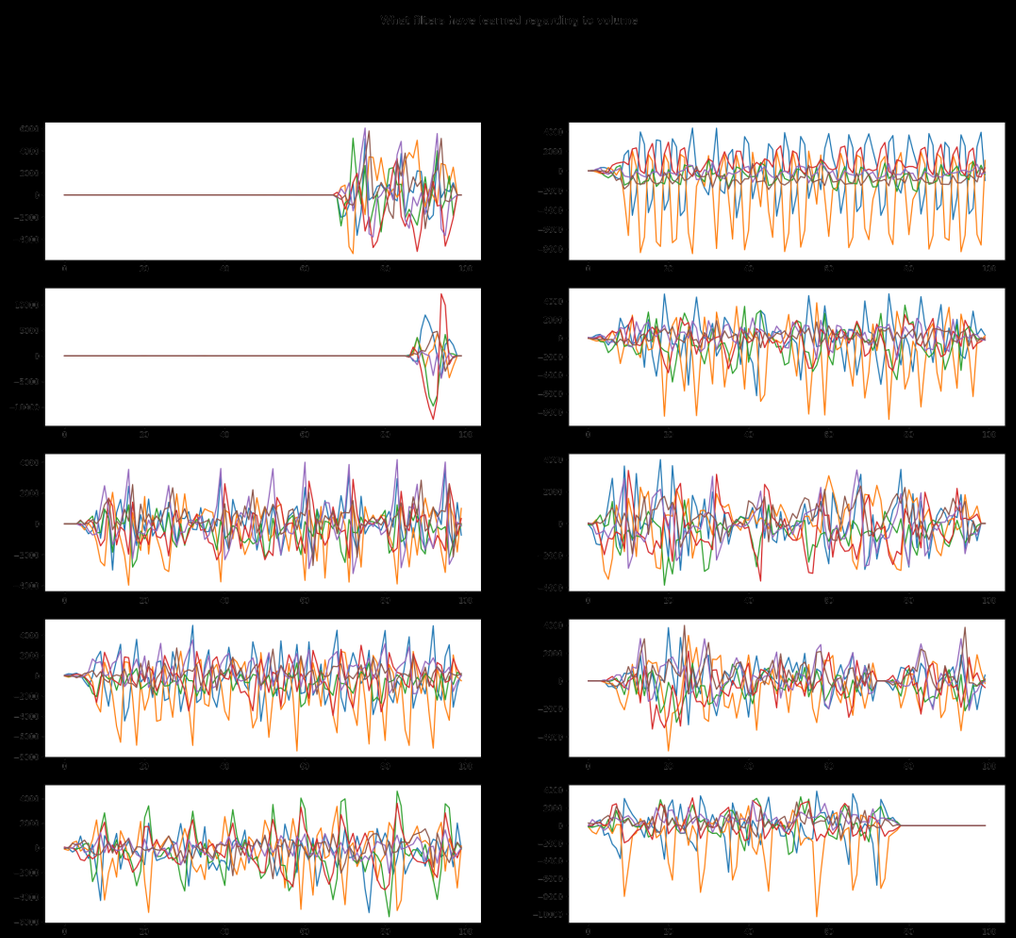
模型给“中价将会上升”分类评分最高的数据



模型给“中价将会下降”分类评分最高的数据

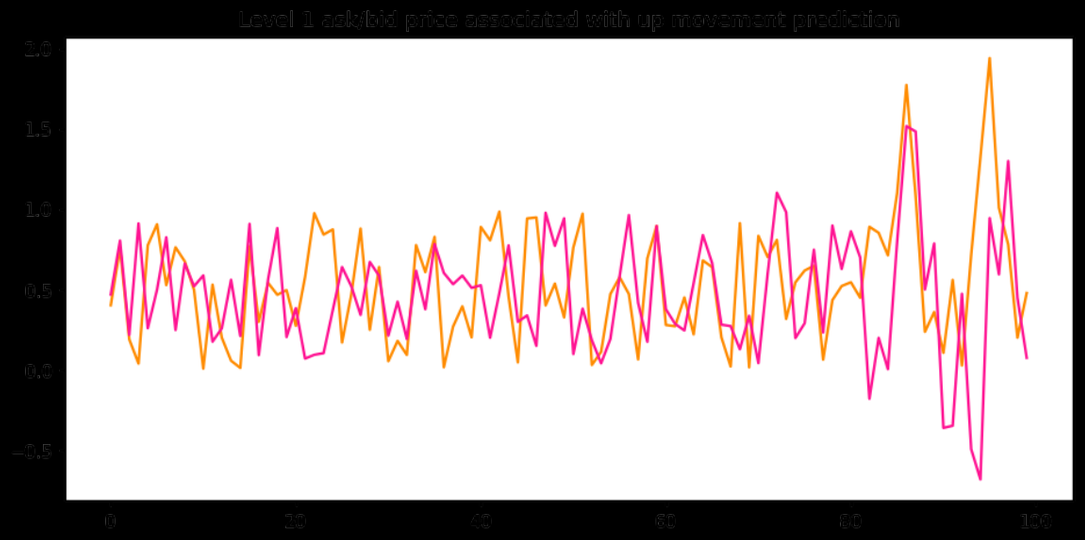
CNN中某个靠后的激活层的滤波器学到了什么

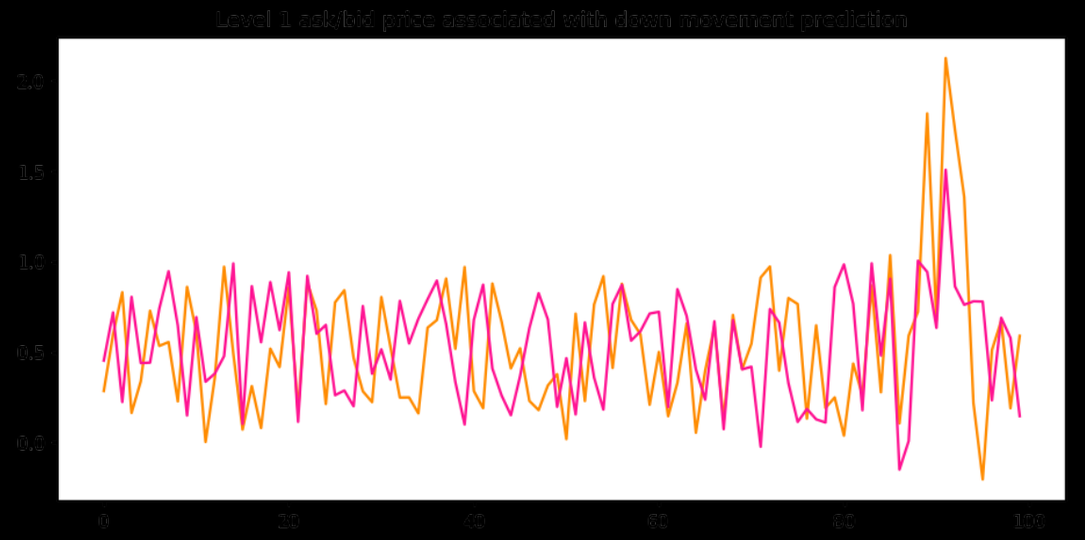




在输入中含有价格的某种变化模式时，这一层会积极地作出反应

CNN最后的全连接层学到了什么





- 窗口尾部出现了方向相反的较高的峰——买卖价格反弹

限制

- 空洞卷积模型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
- 可视化时间序列的数据方法不如T–SNE等方法好