## 基于VMD-Attention-LSTM模型收盘价预测深度学习模型

### 数据处理

1.基本数据处理

原数据分为全国七个碳交易试点日交易价格数据（北京、上海、深圳、湖北、广东、天降、重庆、福建）与全国数据，因为福建省的数据时间与其他数据时间差距较大所以舍弃，全国数据过于笼统，不符合我们预测湖北收盘价的需求，我们也将其舍弃，并取出剩下各省份的收盘价整理出一份《处理后的数据表.CSV》。

因为使用的模型主要为Attention-LSTM，参数设计较大，我们需要更多的数据量，我们在原数据处理上，我们以1135条数据量的北京市收盘价为基准特征，将北京、深圳、广东、上海、湖北的数据组成输入VMD分解的特征（1135组\*5个每组一个时间步，共5675个数据），此步骤主要是避免因为其他省份数据组合后数据过少与选取的数据中出现Nan值。

#### VMD特征分解

在utilt目录中设计VMD类与VMD分解后特征对原数据 点对应处理函数vmd\_sp\_lis、vmd\_lis；根据VMD分解的原著 论文设计出VMD特征分解类，在参数设置上经过多次实验， K为3、alpha=2000、tau=0.6时模型效果最佳。

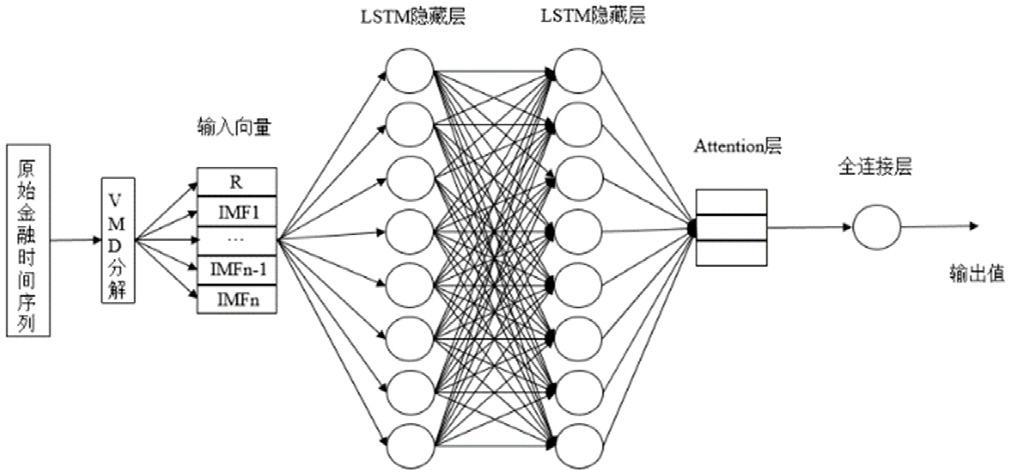
#### 模型训练前的数据处理

根据LSTM层的需求，输入的数据应该为 [送入样本数， 循环核时间展开步数， 每个时间步输入特征个数] ，循环核时间展开步数我在代码中设计为使用前30天的数据，预测出第31天的数据，以此类推，每个时间步的输入特征个数我在代码中的设计为将当天的分解后特征每个时间段仅对应其中五个原数据的VMD的分解特征，经过实验当以所有时间原数据分解后的特征作为特征输入进网络时，数据量会被压缩的过小，导致网络严重的过拟合问题，当我们输入当前对应其五个原数据的VMD的分解特征时，对下推移并不会对结果产生影响，且模型过拟合问题得到巨大的缓解。

最后将数据分为训练集、测试集，对应的形状为（840，30，15）、（205，30，15）送入网络即可。

### 模型设计

模型设计在models下的vmd\_attention\_lstm下，模型分为两个部分，Attention层attention\_3d\_block 为点积注意力模块，Attention\_lstm则为最终模型结构；其中模型结构中使用了两个128单元的LSTM层，一层Attention\_LSTM组合，一层展平层及一层全连接输出层，并为了防止过拟合使用了Dropout层其参数为0.5，已经尽可能不影响输出的正则化函数；其中输入为（送入样本数，时间步，VMD分解后的特征），以此构成VMD-Attention-LSTM模型，模型结构图如下：

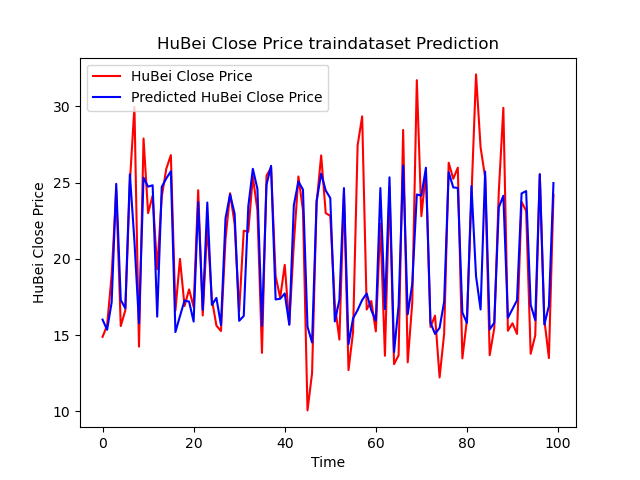


### 模型训练

模型训练上，经过测试128个LSTM神经元数量是个好的选择，学习率使用1e-4，Batch Size为128，使用CallBack函数返回其最优模型权重参数，使用Adam优化器及Huber损失函数，因为数据量小，我们使用500次训练迭代次数，取得其中最好的模型权重。

### 模型结果

最后，设计出预测应用，读取模型及保存后的权重信息；获取湖北原数据对数据进行预处理后的后100个时间点的数据输入预测模型，最后得到模型预测结果图：



可见数据量小的情况下，经过一系列的模型调整，我们可以得到一个相较吻合的预测数据；评价后的数据为：

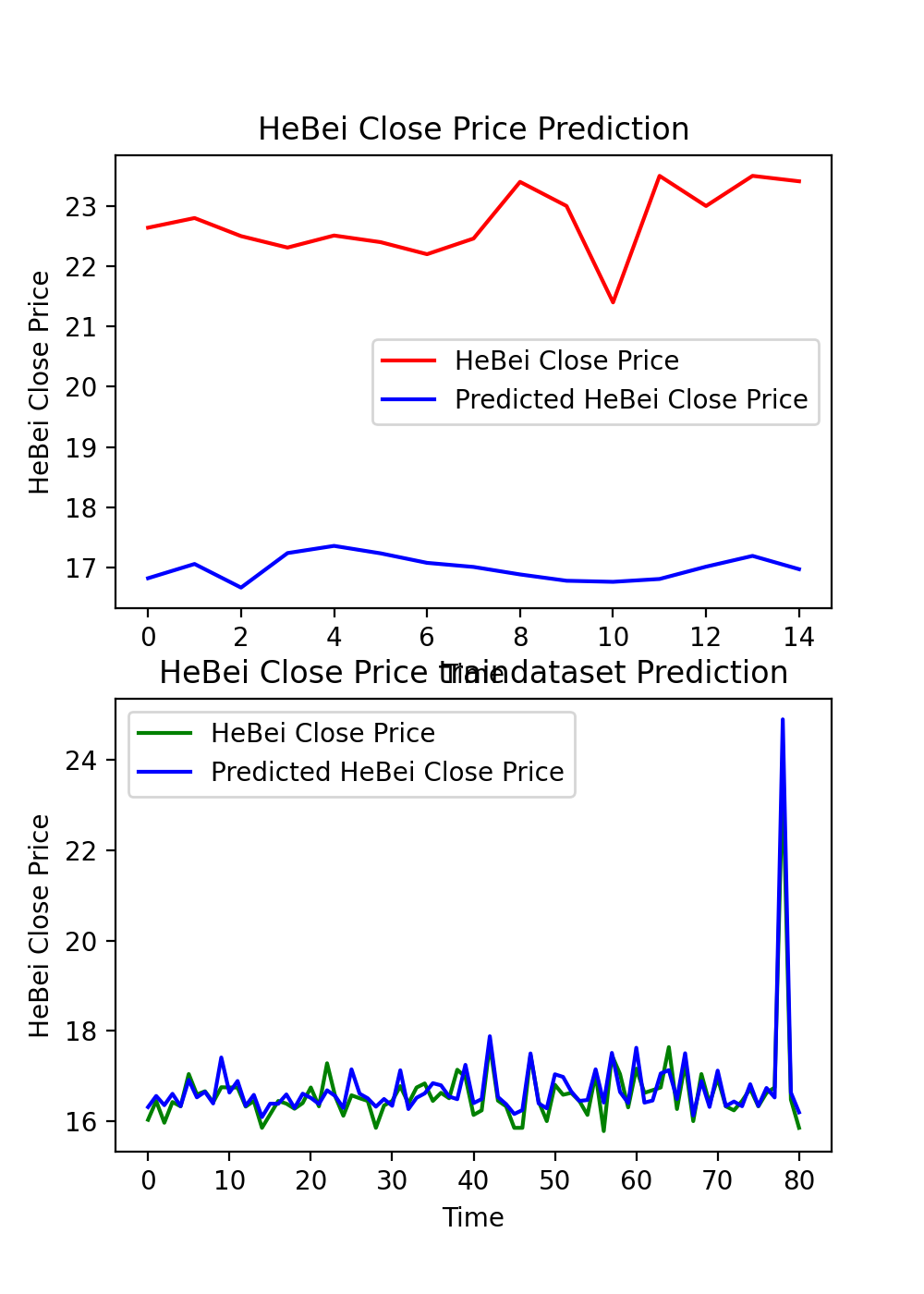
MSE: 9.909498

RMSE: 3.147936

MAE: 1.912804

MAPE: 0.103443

最后附上模型调整完毕且良好训练前，我的遇到的过拟合时的预测与真实的对比图（上为与验证数据的拟合图，下为与训练数据的拟合图）：



可见其明显的存在过拟合问题。

思考：通过一系列防止过拟合处理，最后还是无法改善过拟合问题，最后只能通过数据改进的方式去改善其问题，其实如果数据量够大，以三个分解量\*五个特征值\*30个时间段的特征组合将会达到更好的模型效果，当时以12个分解量\*五个特征值\*5个时间段的特征组合训练时训练Loss一度接近0的值也可以印证我的猜想。