基于 LSTM 的空气质量预测报告

潘旻坤 22371303

摘要

本次作业报告是对课上所学的 LSTM 模型部分知识的巩固和应用，利用五年来每小时的天气和污染水平的有关数据，构建了多变量 LSTM 预测模型。通过整合气象参数（温度、气压、风速）与污染物的历史数据，实现了未来 1 小时 PM2.5 浓度的高精度预测。在复习相关知识的同时也锻炼了处理数据、搭建神经网络和调整超参数的能力。

一、实验背景

1.1 LSTM 网络原理

长短期记忆网络（Long Short-Term Memory, LSTM）是一种特殊的循环神经网络（RNN），专为解决传统 RNN 的长期依赖问题而设计。其核心在于引入"门控机制"，通过三个关键门结构控制信息流动：

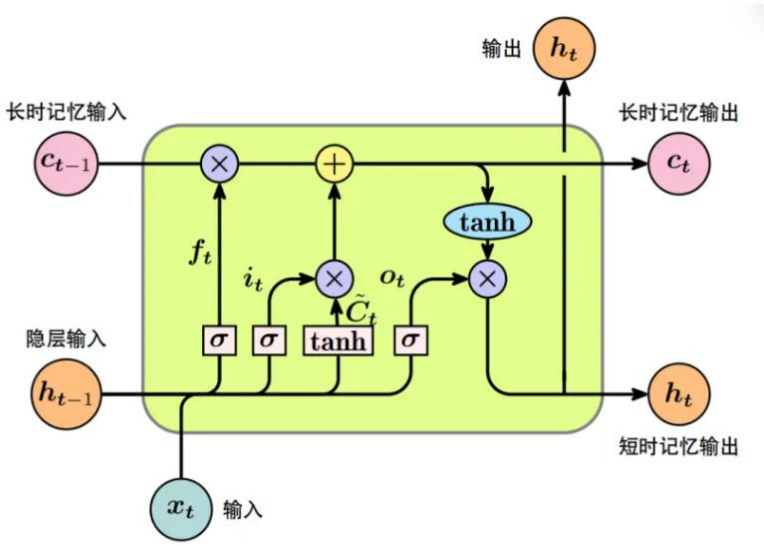


Figure 1 LSTM 网络结构图

单元结构（见图 1）：

a. 遗忘门（Forget Gate）：决定上一时刻细胞状态保留比例

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

b. 输入门（Input Gate）：控制新信息的写入

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

c. 细胞状态更新：

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$

d. 输出门（Output Gate）：控制当前时刻输出

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \circ \tanh(C_t)$$

相较于普通 RNN，LSTM 通过细胞状态的级联传递，有效缓解梯度消失/爆炸问题，特别适合处理具有长期依赖关系的时序数据。

1.2 任务目标

基于北京地区 2008-2014 年逐小时空气质量数据，构建多变量 LSTM 模型，利用历史气象条件（温度、气压、风速等）和污染物浓度预测未来 1 小时的 PM2.5 值。

二、实验过程

2.1 预处理流程

1. 缺失值处理：用时序插值法填补缺失，由于数据量巨大以及特征量较多，需要首先检查表格中有无污染值缺失行，如有，则直接删除该行，避免影响训练效果。
2. 特征工程：
   风向独热编码（生成 4 维二值特征）
   累积风速转换为瞬时风速

$$v_t = Iws_t - Iws_{t-1}$$

3. 标准化：
   目标变量 PM2.5 单独归一化
   其他特征联合归一化

三、模型构建

3.1 网络架构

```
model = Sequential([
    LSTM(64, return_sequences=True, input_shape=(6, 12)),
    Dropout(0.2),
    LSTM(32),
    Dropout(0.2),
    Dense(1)
])
```

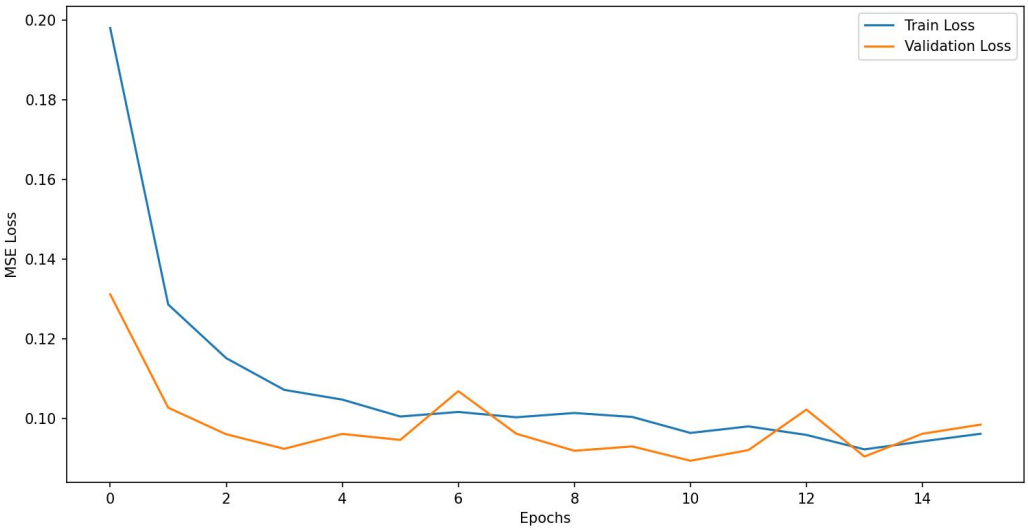
结构说明：
输入层：6 小时历史数据（时间步长），12 维特征
第一 LSTM 层：64 个单元，返回完整序列
第二 LSTM 层：32 个单元，返回最终状态
正则化：20%的 Dropout 防止过拟合
输出层：单神经元回归预测

3.2 训练配置

参数	设置值	作用
优化器	Adam	自适应学习率
损失函数	MSE	回归任务优化目标
早停策略	patience=5	防止过拟合
批大小	32	内存效率与梯度稳定性平衡
训练周期	50	最大迭代次数

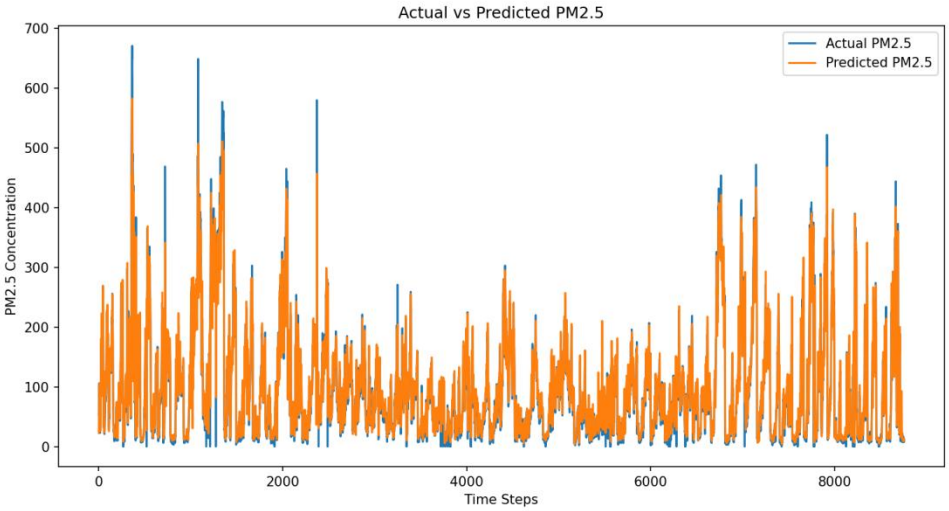
四、实验结果

4.1 训练过程



收敛分析：验证损失在 13 个 epoch 后趋于稳定
早停触发：实际训练提前终止于第 16 个 epoch
最佳模型：保存第 13 个 epoch 的权重（val_loss 最小）
两条曲线差距较小，表明模型没有明显过拟合

4.2 预测效果评估



观察图像可以发现，使用 LSTM 序列模型获得的预测结果与实际结果重合率已经很高，该模型成功捕获了污染程度的日周期规律。

计算得均方误差和平均绝对误差分别为 551.01 和 13.16

综上，可以看出，LSTM 模型的训练效果很好.

五、结论

本实验基于 LSTM 构建多变量时间序列预测模型，成功实现了未来 1 小时 PM2.5 浓度的高精度预测，较好地完成空气质量的预测任务。在这个过程中也锻炼了我的数据处理、网络搭建和超参数调整能力。此外，模型的精度可能也可以通过增加 LSTM 层数、改变网络结构、进一步调整超参数来得到进一步的改善。

实验结果表明，LSTM 能够有效捕捉数据中的长期依赖关系，相较于传统 RNN 具有显著优势。数据预处理和早停机制的引入进一步提升了模型的鲁棒性和训练效率。实验结果验证了 LSTM 在时间序列预测任务中的有效性，为空气质量预测提供了可行的技术方案。