基于 LSTM 的空气质量预测实验报告

钟经佑 22373168

一、引言

随着工业化进程的加速，空气质量问题逐渐成为全球关注的焦点。空气污染不仅对人体健康造成严重威胁，还影响生态环境和社会经济发展。为了有效预测空气质量，科学家们采用了多种机器学习和深度学习方法，其中长短期记忆网络（LSTM）因其在处理时序数据方面的优势，成为解决空气质量预测问题的重要工具。本实验旨在利用 LSTM 模型预测未来的空气质量，基于历史的气象数据和空气污染指数（PM2.5）进行训练与预测。

二、数据集介绍

本实验使用的数据集来源于 Kaggle，该数据集记录了过去五年内每小时的空气质量与气象数据。具体来说，数据集包含以下字段：

pm2.5：PM2.5 浓度（空气质量指数）

DEWP：露点温度（° C）

TEMP：温度（° C）

PRES：气压（hPa）

cbwd：风向的综合编码

Iws：风速（累积风速）

Is：积雪小时数。

Ir：积雨小时数。

数据集分为训练集和测试集，其中训练集包含历史气象数据和污染数据，测试集用于评估模型的预测性能。通过对过去小时的多变量数据进行训练，模型能够预测未来小时的 PM2.5 浓度。

三、模型结构

本实验采用 LSTM（长短期记忆网络）作为核心模型。LSTM 是一种特殊的循环神经网络（RNN），适用于时间序列数据的建模。相比传统的 RNN，LSTM 能够有效解决长期依赖问题，通过门控机制控制信息的传递和遗忘。

LSTM 模型的具体结构如下：

输入层：输入为多维的时序数据，包括过去的 PM2.5 浓度、气象数据等特征。每个时间步的输入大小为 8（即 8 个特征）。

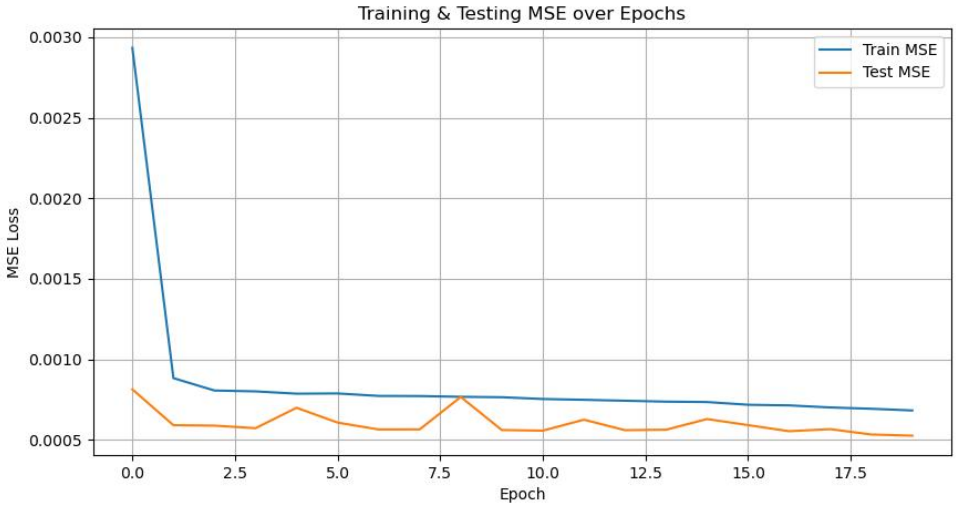
LSTM 层：LSTM 层由多个神经元组成，每个神经元具有一个隐状态和一个细胞状态，用于保存历史信息。模型采用了 2 层堆叠的 LSTM 层，以更好地捕捉数据中的复杂时间序列特征。

全连接层：LSTM 层的输出传入全连接层进行线性变换，将隐藏状态映射到最终的 PM2.5 浓度预测值。

模型的训练目标是最小化均方误差（MSE），即使得预测的 PM2.5 值尽可能接近真实值。

四、训练结果

训练过程中，使用了 24 小时的数据作为输入，预测下一个小时的 PM2.5 浓度。模型经过 30 轮训练，训练集和测试集的 MSE 曲线随着 epoch 的增加逐渐收敛。

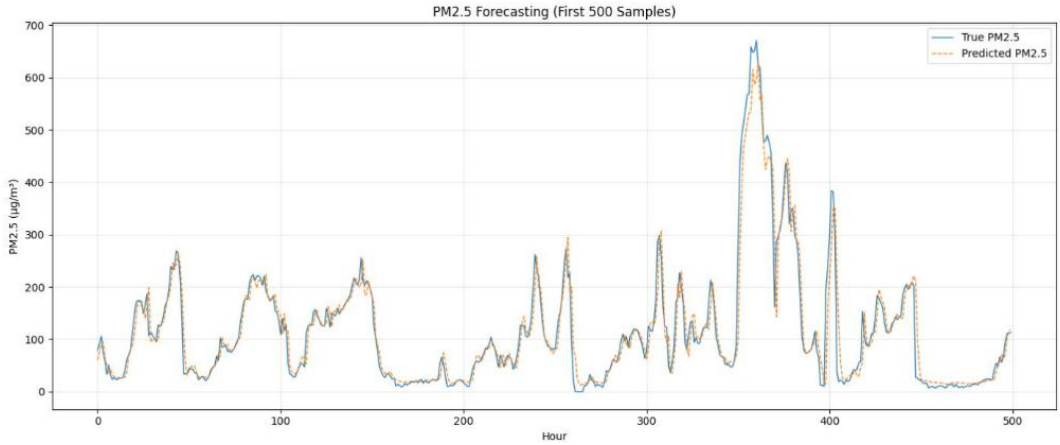


训练过程中的主要观察结果：

训练集 MSE：随着训练轮次的增加，训练集的误差不断减小，表明模型在训练数据上逐渐收敛。

测试集 MSE：测试集的 MSE 也表现出较好的下降趋势，说明模型对未见过的数据具有一定的预测能力。

模型收敛情况：训练过程中，模型的损失函数不断减小，且最终在训练集和测试集上达到了较低的 MSE，这表明模型能够有效地学习到空气质量的时间序列特征。



该图展示了前 500 个时间步的 PM2.5 预测结果，其中横轴表示时间（以小时为单位），纵轴表示 PM2.5 浓度（单位为μg/m³）。图中蓝色线条表示真实的 PM2.5 浓度，而橙色虚线表示模型预测的 PM2.5 浓度。从图中可以看出，尽管模型在某些时刻的预测值与真实值之间存在差异，但整体趋势非常相似。模型能够较好地捕捉到 PM2.5 浓度随时间变化的波动，尤其是在大多数时间段内，预测结果与真实数据吻合得较为紧密。然而，在一些极端变化的时刻，如图中的高峰期，预测误差显著增大，这可能表明模型在处理快速变化的污染水平时存在一定的挑战。总体而言，尽管存在一些误差，模型依然能够较为准确地反映出 PM2.5 浓度的变化趋势，并为未来的空气质量预测提供了有效的参考。

五、总结与展望

本实验成功应用 LSTM 模型进行空气质量预测，利用历史气象数据和空气污染指数来预测未来的 PM2.5 浓度。实验结果表明，LSTM 模型能够较好地捕捉时间序列数据中的长期依赖关系，并且在测试集上表现出较好的泛化能力。

然而，本实验也存在一些不足之处，例如：

数据质量问题：尽管数据经过了预处理和归一化，但仍然可能存在噪声数据或异常值，未来可以考虑更为精细的数据清洗方法。

模型优化：目前使用的 LSTM 模型结构较为简单，未来可以尝试不同的 LSTM 变体（如双向 LSTM、Attention LSTM）来进一步提升预测性能。

多模态数据融合：除了气象数据，其他因素如交通流量、社会活动等也可能对空气质量产生影响，未来可以结合更多外部特征进行模型的改进。

总之，基于 LSTM 的空气质量预测是一个具有前景的研究方向，未来可以在多领域的应用中进一步推广，包括城市空气污染预测、环境监测等。