基于 LSTM 模型的预测任务实验报告

王雅妮 wangyani@buaa.edu.cn

摘要

本实验利用长短期记忆网络（LSTM）模型，对空气污染数据进行预测。通过对包含气象信息和历史污染数据的时间序列进行建模，实现了对未来空气污染程度的有效预测。实验结果表明，该模型能够较好地学习空气污染的变化规律，并具备一定的预测能力。

引言

传统的时间序列分析方法在处理复杂的非线性时间序列时存在局限性，而深度学习模型，尤其是 LSTM，在处理时间序列数据方面展现出强大的能力。本研究旨在利用 LSTM 模型对空气污染数据进行预测，并评估其性能。

方法原理

LSTM 是一种特殊的循环神经网络（RNN），通过引入门控机制，有效地解决了传统 RNN 在处理长序列时出现的梯度消失和梯度爆炸问题。LSTM 能够更好地捕捉时间序列中的长期依赖关系，因此在时间序列预测任务中表现出色。

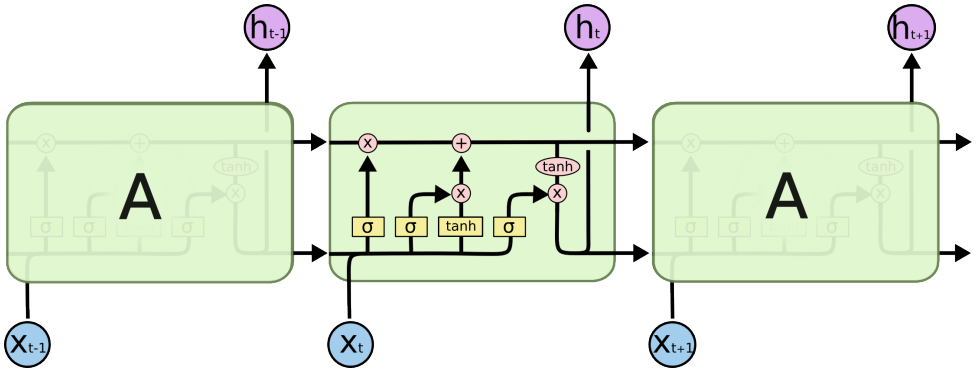
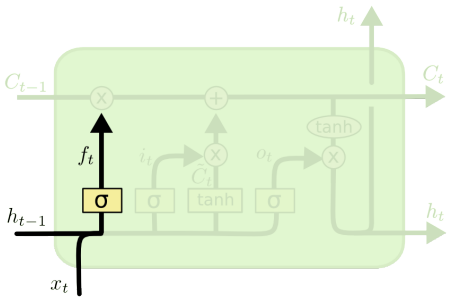


Figure 1: LSTM 内部结构示意图

LSTM 单元的核心是细胞状态（cell state），它类似于一个信息传送带，在整个序列中传递信息。门控机制负责控制信息的流入、流出和遗忘，从而使 LSTM 能够选择性

地记忆和遗忘信息。

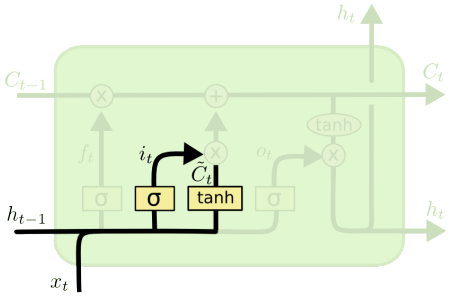
遗忘门由一个 *sigmiod* 神经网络层和一个按位乘操作构成，它决定了细胞状态 C_{t-1} 中的哪些信息将被遗忘。



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

Figure 2: 遗忘门

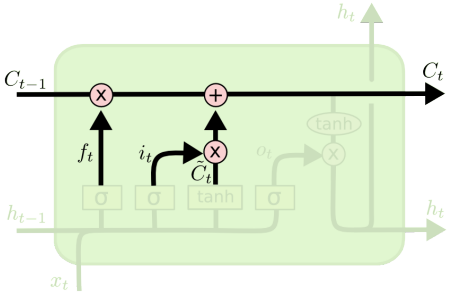
记忆门由输入门（input gate），即一个 *sigmoid* 神经网络层，与 *tanh* 神经网络层和一个按位乘操作构成。记忆门的作用与遗忘门相反，它将决定新输入的信息 x_t 和 h_{t-1} 中哪些信息将被保留。



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

Figure 3: 记忆门

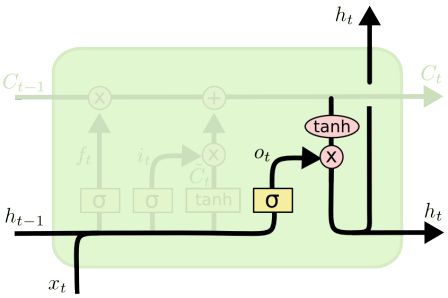
由遗忘门和记忆门的输出，可以更新细胞状态 C_t。C_t 将继续传递到 $t+1$ 时刻的 LSTM 网络中，作为新的细胞状态传递下去。



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

Figure 4: 更新细胞状态

输出门（output gate）即一个 *sigmoid* 神经网络层，与 *tanh* 函数以及按位乘操作共同作用将细胞状态和输入信号传递到输出端。



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

Figure 5: 输出门

实验方法

1. 数据集

本研究使用的数据集包含气象信息（如温度、湿度、风速等）和历史空气污染数据。数据集被分为训练集和测试集，用于模型的训练和性能评估。

2. 数据预处理

（1）风向数据处理：对风向数据进行独热编码，将文字信息转换为数值型特征。

（2）特征合并：将气象数据和独热编码后的风向数据合并成特征矩阵。

（3）数据归一化：使用 MinMaxScaler 对特征矩阵和目标变量（空气污染程度）进行归一化，将其缩放到 [0, 1] 范围内。

3. 模型构建

本研究采用的 LSTM 模型结构如下：

（1）LSTM 层：包含 64 个单元。

（2）Dropout 层：Dropout 率为 0.2。

（3）Dense 层：包含 32 个单元，激活函数为 ReLU。

（4）Dense 层：输出层，包含 1 个单元。

4. 模型训练

（1）时间窗口：使用时间窗口为 12 的数据构建训练样本，即用过去 12 小时的数据预测下一小时的污染程度。

（2）优化器：使用 Adam 优化器，学习率为 0.001，并设置梯度裁剪（clipvalue=1.0）。

（3）损失函数：使用均方误差（MSE）。

（4）学习率调度器：使用 ReduceLROnPlateau，监控验证集损失，当损失不再下降时降低学习率。

（5）早停机制：使用 EarlyStopping，监控验证集损失，当损失连续 20 个 epoch 没有改善时提前停止训练。

（6）训练过程：将训练集进一步划分为训练集和验证集（8:2），训练模型 100 个 epoch。

实验结果

1. 训练过程

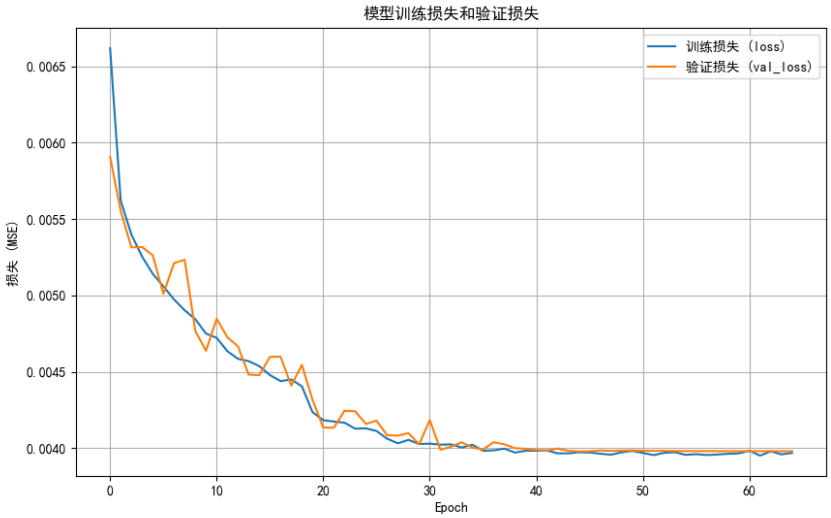


Figure 6: 训练损失和验证损失随 epoch 变化曲线

训练损失和验证损失均随着 epoch 的增加而下降，最终收敛到一稳定值，在早停机制作用下提前结束训练。模型能够有效地学习训练数据，在验证集上亦有较好表现。

2. 预测结果

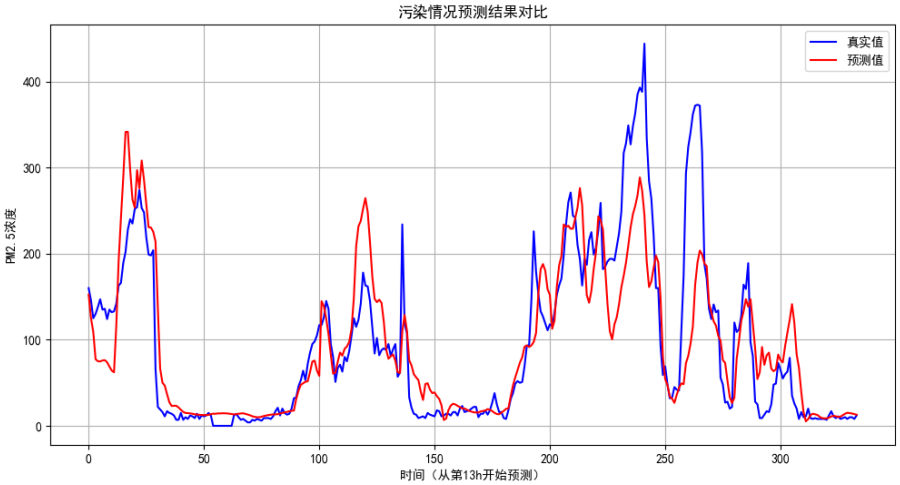


Figure 7: 测试集上真实值与预测值的对比曲线图

对比曲线表明，模型能够较好地预测空气污染程度的变化趋势，但可能在某些峰值处存在一定的预测误差。此外，由于 LSTM 需要前 $12h$ 的数据输入用于预测，本实验的预测结果实际上是从第 $13h$ 开始的。

测试集上的均方误差（MSE）为 3183.8833.

结论

本研究利用 LSTM 模型对空气污染数据进行了预测，并取得了较好的结果。实验表明，LSTM 模型能够有效地捕捉空气污染时间序列中的复杂模式，并具备一定的预测能力。然而，模型在预测峰值方面仍有提升空间，未来的工作可以考虑尝试更复杂的模型结构，以进一步提高预测精度。