Software Engineering and Applications 软件工程与应用, 2020, 9(2), 135-142
Published Online April 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.92016

Hans汉斯

Improved LSTM-Based Ozone Concentration Prediction Model Based on Time Series Delay Correlation Algorithm

Zhixin Tie[1,2], Xiaoning Cheng[2], Deshou Lin[2], Chengfu Ding[3]

[1]School of Science and Technology, Zhejiang University of Technology, Shaoxing Zhejiang
[2]Zhejiang University of Technology, Hangzhou Zhejiang
[3]Hangzhou Juguang Technology (Hangzhou) Co., Ltd., Hangzhou Zhejiang
Email: chengxnxn@qq.com

Received: Apr. 7th, 2020; accepted: Apr. 21st, 2020; published: Apr. 28th, 2020

Abstract

Ozone pollution has attracted increasing attention. How to accurately predict ozone concentration has become an important subject. Taking the hourly monitoring data of ozone at multiple sites in Hangzhou as the research object, the correlation of ozone concentration changes at different sites was analyzed, and a long-term short-term memory (LSTM) neural network model was proposed to improve the LSTM ozone concentration prediction based on a time series delay correlation algorithm. The model is compared with the traditional LSTM model and SpaceLSTM model, and the results show that the proposed method has the smallest mean square error and the prediction result is more accurate.

Keywords

Ozone, LSTM, Time Series Delay Correlation Algorithm

基于时间序列延迟相关算法改进LSTM的臭氧浓度预测模型

铁治欣[1,2]，程晓宁[2]，林德守[2]，丁成富[3]

[1]浙江理工大学科技与艺术学院，浙江 绍兴
[2]浙江理工大学，浙江 杭州
[3]聚光科技(杭州)股份有限公司，浙江 杭州
Email: chengxnxn@qq.com

收稿日期：2020年4月7日；录用日期：2020年4月21日；发布日期：2020年4月28日

文章引用: 铁治欣, 程晓宁, 林德守, 丁成富. 基于时间序列延迟相关算法改进 LSTM 的臭氧浓度预测模型[J]. 软件工程与应用, 2020, 9(2): 135-142. DOI: 10.12677/sea.2020.92016

摘　要

臭氧污染日益引起人们的重视，如何准确预报臭氧浓度成为一个重要课题。以杭州市多个站点臭氧的小时监测数据为研究对象，分析不同站点的臭氧浓度变化的相关性，结合长短期记忆(LSTM)神经网络模型提出一个基于时间序列延迟相关算法改进LSTM的臭氧浓度预测模型，通过与传统的LSTM模型和SpaceLSTM模型进行对比实验，结果表明：所提出的方法均方误差最小，预测结果更加准确。

关键词

臭氧，LSTM，时间序列延迟相关算法

1. 引言

臭氧本身具有独特的鱼腥味[1]，这也是最初被人们发现的主要原因。当少量臭氧存在于人类生活的对流层时，对人类生活和健康状况不会产生威胁[2]。高浓度的臭氧环境会损害人类的心肺功能，诱发呼吸道等各类疾病。在对珠江三角洲地区的研究表明，臭氧浓度与每日的死亡率呈现正相关性，对急性死亡具有明显的影响[3]。城市中的臭氧主要由氮氧化物(NOx)、挥发性有机物(VOCs)在适宜条件下作用产生[4]，随着现代工业的发展，煤矿、石油的燃烧以及汽车尾气的大量排放，空气中的臭氧浓度严重超标，2015年中国环保部发布的京津冀长三角、珠三角区域以及74个城市空气质量状况的报告显示，臭氧已经成为影响空气质量的首要污染物[5]。严峻的臭氧环境污染使得对臭氧浓度预测的研究不容忽视，尤其是大数据时代的到来，更多的数据分析被运用到臭氧浓度预测上，Alqamah Sayeed用深层卷积神经网络开发一个提前24小时预测臭氧浓度的模型，证实了模型的可接受准确性，但同时也提出，该模型的预测结果仍低估了每日的臭氧含量，准确率不够，为以后的研究和改进提供方向[6]。王振友提出运用矩阵改进GM模型对大气中的臭氧含量进行分析，使得预测结果相对误差在6%以内，与实际数据吻合度良好[7]；朱佳运用小波分解分离复杂的信号频率，结合最小二乘支持向量机建立臭氧预测模型，并通过对比实验，证明了模型优于SVM模型和ANN模型[8]；张春露在针对SVM模型、ARIMA模型以及LSTM模型在AQI指数上预测的有效性问题进行了对比实验，得出LSTM模型预测精度最高的实验结果[9]。为了更好的对臭氧浓度进行预测，本文运用时间序列延迟相关算法对LSTM模型进行改进，提出了基于时间序列延迟相关算法改进的LSTM模型(TD-LSTM)，能够有效的提升臭氧浓度的预测精度，在相同的条件下，运用杭州市多个站点监测的臭氧浓度数据进行对比实验，证实了提出的预测模型的有效性。

2. 相关技术介绍

2.1. 时间序列延迟相关算法

时间序列延迟相关算法是时间序列数据挖掘的重要研究内容，目前已经在股票市场、气候分析、天气预报等领域得到应用[10]。具体来说，对于两个时间序列 $A=\{X_t \mid t=0,1,2,\cdots,n-1\}$ 和 $B=\{y_t \mid t=0,1,2,\cdots,n-1\}$ 进行计算分析，找到两个序列延迟相关性最大时的延迟时间[11]，延迟相关是指两个时间序列的最大相似

度不是发生在 $t=0$ 的时刻，而是 $t=s\left(s\neq 0\right)$ 的时刻，此时 s 就是延迟的大小，计算公式如式(1)、(2)所示：

$$R\left(s\right)=\frac{\sum_{t=s}^{n-1}\left(x_t-\overline{x}\right)\left(y_{t-s}-\overline{y}\right)}{\sqrt{\sum_{t=s}^{n-1}\left(x_t-\overline{x}\right)^2}\sqrt{\sum_{t=0}^{n-s-1}\left(y_t-\overline{y}\right)^2}} \tag{1}$$

$$\overline{x}=\frac{1}{n-s}\sum_{n=s}^{n-1}x_t,\overline{y}=\frac{1}{n-s}\sum_{t=0}^{n-s-1}y_t \tag{2}$$

s 的变化区间为 $\left[0,n/2\right]$，当延迟的位置 s 较大时，实验的误差也会变大。当 s 的值从 0 变化到 $n/2$ 时，会产生多个 $R\left(s\right)$ 的值， $R\left(s\right)$ 最大的值就是两个序列相关性最大的时候，称为“最大延迟相关点”，此时的延迟 s 就是需要的最优值。

2.2. LSTM 神经网络

长短期记忆模型(Long Short-Term Memory neural network, LSTM)是一种特殊的递归神经网络[12]，和传统的递归神经网络(Recurrent Neural Networks, RNN) [13]有所不同，将隐藏层更换成LSTM模型的细胞单元，使其可以长期记忆，便于处理长时间延迟的时间序列。它是一种基于时间序列的模型，它的选择性记忆门能够更好地建立先前的信息和当前环境的时间相关性，是对 RNN 模型的一种改进。LSTM 神经网络有三个门控，输入门、输出门、遗忘门，其模块示意图如图 1 所示，矩形表示神经网络层，圆形表示逐点操作。控制门主要是由一个 sigmoid 函数[14]与点乘操作组成，可以决定多少信息可以传送出去，记忆门中增加存储单元来存储历史信息，可以有效解决神经网络中的梯度消失问题，可以更加深入挖掘时间序列中存在的规律。

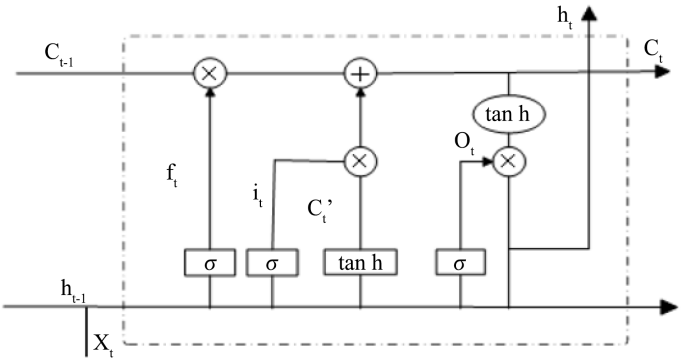

Figure 1. LSTM gating model diagram
图 1. LSTM 门控模型图

假设 f_t,i_t,o_t 分别表示在 t 时刻，遗忘门、输入门和输出门的数值，则：

$$i_t=\sigma\left(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}C_{t-1}+b_i\right) \tag{3}$$

$$f_t=\sigma\left(W_{xf}x_t+W_{hf}h_{t-1}+W_{cf}C_{t-1}+b_f\right) \tag{4}$$

$$o_t=\sigma\left(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}C_t+b_o\right) \tag{5}$$

其中， x_t 表示t时刻输入的数据， h_{t-1} 表示 $t-1$时刻的输出值， C_{t-1} 表示 $t-1$时刻的单元记忆值， W_{**} 表示权重系数， b_* 表示偏置向量， σ 表示sigmoid函数，值为0~1之间，当为0时，当前信息不传送，当为1时，当前信息全部传送。

3. 模型介绍

3.1. 基于空间分布的 LSTM 预测模型(SpaceLSTM)

基于空间的 LSTM 模型(SpaceLSTM)，将污染物空间传播的特性考虑在内，在原始 LSTM 模型的基础上，利用不同空间分布上多个监测站点的臭氧浓度的历史数据，对当前时刻的臭氧浓度进行预测。该模型输入多个不同地理位置站点的臭氧浓度数据，经过数据预处理之后经由 LSTM 模型，得出臭氧浓度的预测结果。实验结果运用均方根误差(RMSE)作为评价指标，具体公式如(6)所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-X_r\right)^2} \tag{6}$$

其中，X_i 表示臭氧浓度的预测值，X_r 表示臭氧浓度的真实值，n 表示训练数据集的样本数量。均方根误差值越小，表明模型预测的准确度越高。

3.2. 基于时间序列延迟相关算法改进的 LSTM 模型(TD-LSTM)

针对基于空间分布的 LSTM 模型(SpaceLSTM)忽略了空间分布上不同站点之间臭氧污染因子影响不同的问题，提出基于时间序列延迟相关算法改进的 LSTM 模型(TD-LSTM)，运用时间序列延迟相关算法，考虑不同站点数据之间的延迟相关性，使得预测模型更加准确。

TD-LSTM 模型的流程图如图 2 所示，具体执行步骤分为以下四步：

1) 数据预处理。由于数据样本是由多个不同站点数据组成，不同站点之间使用不同的量纲和量级，并且数据因子范围较大，为了实验的准确性和高效性，需要对实验数据进行归一化的处理，本文采用 min-max 标准化法，计算公式如(7)所示：

$$x_t'=\left(x_t-x_{\min}\right)/\left(x_{\max}-x_{\min}\right) \tag{7}$$

其中，$x_{\max}$、$x_{\min}$ 分别表示空气质量检测数据中的最大值和最小值，x_t' 表示归一化之后对应的监测数值，x_t' 为实际监测值。

2) 运用时间序列延迟相关算法，计算数据输入矩阵。

将臭氧浓度小时监测数据按监测时间的先后转换成时间序列，每个站点数据对应一条时间序列，用 $x_t\left(t=1,2,\cdots,n\right)$ 表示需要预测的目标站点 S 在 t 时刻的监测数据，$x_t^j\left(t=1,2,\cdots,n;\ j=1,2,\cdots,m\right)$ 表示站点 S 的编号为 j 的周边站点在 t 时刻的监测数据，其中，m 表示周边监测站点的个数，则原始输入样本数据矩阵为 T_o，如公式(8)所示：

$$T_o=\begin{bmatrix}X\\X^1\\X^2\\\vdots\\X^m\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\\x_1^1 & x_2^1 & \cdots & x_n^1\\x_1^2 & x_2^2 & \cdots & x_n^2\\\vdots & \vdots & \ddots & \vdots\\x_1^m & x_2^m & \cdots & x_n^m\end{bmatrix} \tag{8}$$

其中 $X=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\end{bmatrix}$，$X^j=\begin{bmatrix}x_1^j & x_2^j & \cdots & x_n^j\end{bmatrix}$。

假设周边站点 $j\left(j=1,2,\cdots,m\right)$ 与目标站点 S 达到最大延迟相关时，它与目标站点 S 的最大延迟相关点为 s_j，由于延迟的概念是相对的，若时间序列 A 延迟于时间序列 B 的时间为 s_j，我们也可以认为时间序列 B 超前时间序列 A 的时间为 s_j 则输入样本数据矩阵变为 T_m，如公式(9)所示：

$$T_m = \begin{bmatrix} X \\ TX^1 \\ TX^2 \\ \vdots \\ TX^m \end{bmatrix} \tag{9}$$

其中，当周边站点 $j(j=1,2,\cdots,m)$ 延迟于目标站点 S 时，$TX^j = \begin{bmatrix} x_{s_j+1}^j & x_{s_j+2}^j & \cdots & x_n^j & \underbrace{0 \quad 0 \quad \cdots \quad 0}_{s_j} \end{bmatrix}$，当周边站点 $j(j=1,2,\cdots,m)$ 超前于目标站点 S 时，$TX^j = \begin{bmatrix} \underbrace{0 \quad 0 \quad \cdots \quad 0}_{s_j} & x_1^j & x_2^j & \cdots & x_{n-s_j}^j \end{bmatrix}$ 令所有的延迟的最大值为 s_D，所有的超前最大值为 s_B，则得出经由时间序列延迟相关算法筛选得到的输入样本数据矩阵 T_i，如公式(10)所示。

$$T_i = \begin{bmatrix} x_{s_B+1} & x_{s_B+2} & \cdots & x_{n-s_D} \\ x_{s_B+1}^1 & x_{s_B+2}^1 & \cdots & x_{n-s_D}^1 \\ \vdots & \vdots & \ddots & \vdots \\ x_{s_B+1}^m & x_{s_B+2}^m & \cdots & x_{n-s_D}^m \end{bmatrix} \tag{10}$$

3) 将数据划分成训练集与测试集，用训练集数据进行模型训练，将训练集数据处理后的输入样本数据矩阵 T_i 输入 LSTM 模型，并对实验结果进行记录与分析。

4) 用测试数据集对模型训练成果进行验证。

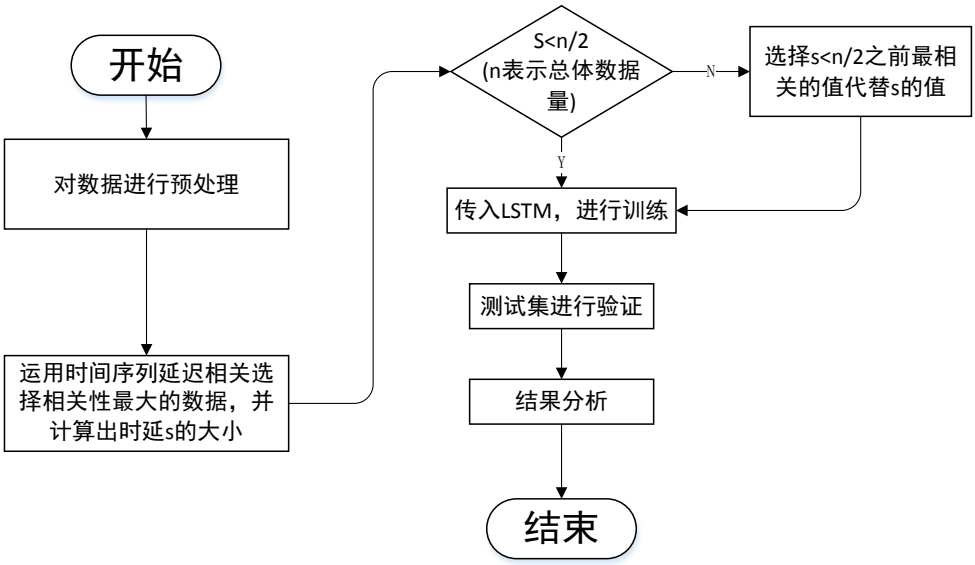

Figure 2. TD-LSTM model flow chart
图 2. TD-LSTM 模型流程图

3.3. 实验数据选取

本文采用杭州市不同站点的空气质量监测数据作为研究对象，实验采取 2019 年 5 月的杭州市各个地区多个站点的臭氧浓度小时监测数据作为实验对象，部分数据样本如表 1 与表 2 所示：

Table 1. Hourly data information table of ozone concentration monitored by the station
表 1. 站点监测的臭氧浓度小时数据信息表

时间数据 站点编号	1229A	1223A	1224A	1226A	1227A	1228A	1230A	1231A	1232A	1233A
2019-05-31 23:00	61	63	27	88	40	53	63	90	38	38
2019-05-31 22:00	136	88	58	105	54	107	70	124	63	66
2019-05-31 21:00	146	162	141	114	64	148	130	163	102	73
2019-05-31 20:00	193	202	184	139	66	179	170	195	157	95
2019-05-31 19:00	203	213	185	202	122	197	205	219	191	118
2019-05-31 18:00	190	205	205	211	181	192	190	210	188	189

Table 2. Part of 1229A site monitoring data
表 2. 部分 1229A 站点的监测数据

时间数据 污染因子	$Pm_{2.5}$ ($\mu g/m^3$)	Pm_{10} ($\mu g/m^3$)	O_3 (mg/m^3)	NO_2 ($\mu g/m^3$)	CO ($\mu g/m^3$)
2019-05-31 23:00	74	105	61	118	1
2019-05-31 22:00	69	90	136	69	1
2019-05-31 21:00	67	88	146	62	1
2019-05-31 20:00	63	80	193	42	0.9
2019-05-31 19:00	56	70	203	37	0.8
2019-05-31 18:00	29	36	190	24	0.6

3.4. 实验结果分析

为了验证本文提出的预测模型的有效性，选取了传统的 LSTM 模型、SpaceLSTM 模型等进行对比实验，三个预测模型都在相同的实验平台和环境下进行实验。本次对每个模型都进行了 50 次的实验，实验迭代 50 次，每次迭代结束之后计算均方误差，选取实验结果的平均值作为本文的公布数据，并且画出实验结果图。

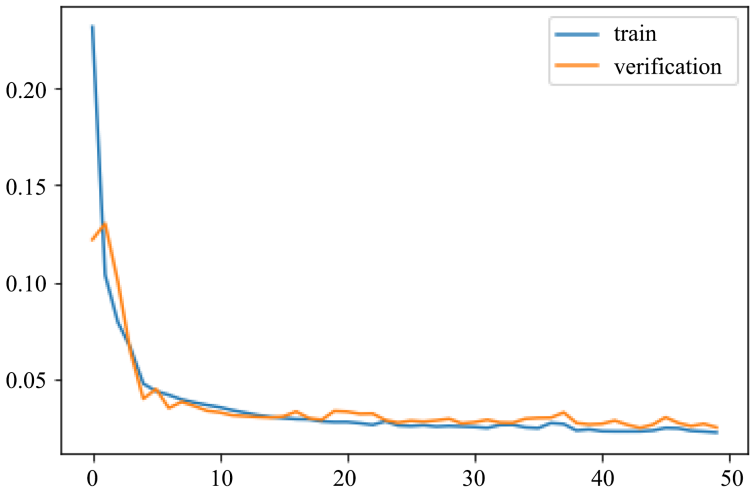

Figure 3. LSTM loss curve after improved delay
图 3. 时延改进后的 LSTM 损失曲线图

图 3 是本文提出的 TD-LSTM 模型在训练过程中，训练集和测试集每一步训练结束时训练数据的损失，横坐标表示迭代次数，纵坐标表示损失值，表示对单个样本而言模型的准确程度，损失值越低，表明准确程度越高，可以看出，迭代超过 10 次以后，模型的损失值可以迅速降低，预测的后期，模型的图像出现上下波动，说明训练已经达到饱和状态，模型基本稳定。

图 4 是采用 TD-LSTM 模型的实验过程中，对臭氧浓度的预测值和真实值的对比曲线图，通过曲线的吻合度可以看出，该模型的预测效果比较好，能够很好的跟踪臭氧浓度的变化趋势。根据模型多次实验的结果，可以得出均方根误差的平均值为 12.205，相比未改进前的模型，有较好的提升。

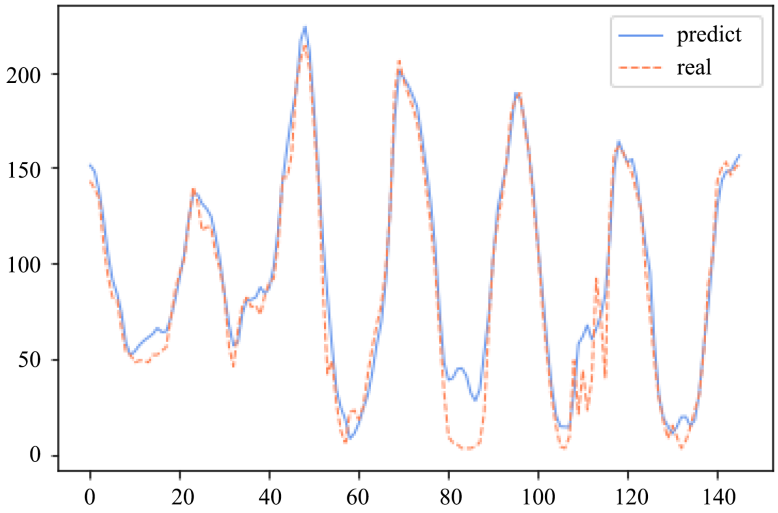

Figure 4. Comparison between predicted and true values
图 4. 预测值和真实值的对比图

如图 5 所示，分别为原始 LSTM 模型(LSTM)、空间数据 LSTM 模型(SpaceLSTM)以及 TD-LSTM 模型的实验对比图，从实验对比图可以看出，时间序列延迟相关改进后的模型与真实值吻合度最高，跟踪臭氧浓度的变化效果也最好。

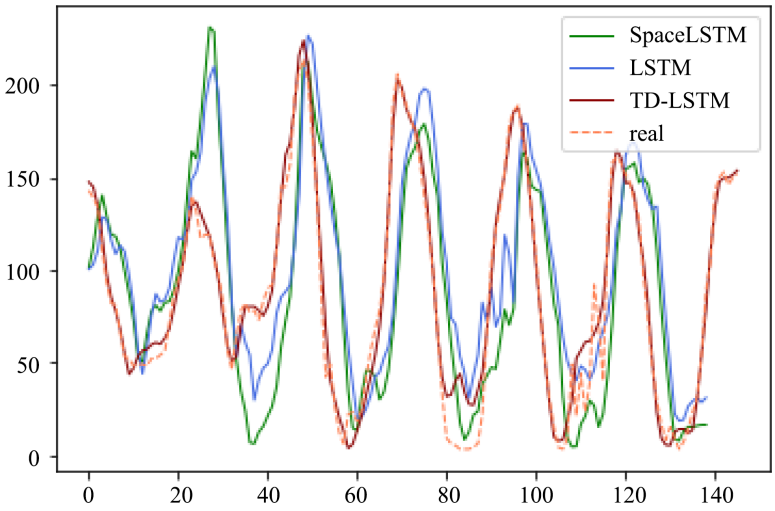

Figure 5. Comparison of various model experiments
图 5. 多种模型实验对比图

从表 3 的数据可以看出，运用时间序列延迟相关算法改进后的 LSTM 模型均方误差值最小，表明 TD-LSTM 模型具有较好的预测精度。

Table 3. Mean mean square error value of ozone (O_3) concentration prediction
表 3. 臭氧(O_3)浓度预测平均均方误差值

模型	LSTM	SpaceLSTM	TD-LSTM
平均均方误差值	21.858	16.392	12.205

4. 结论

本文使用杭州市多个站点臭氧小时监测数据进行臭氧浓度预测。为减少由于量纲造成的误差，首先对臭氧小时监测数据进行 max-min 归一化处理；然后将各站点臭氧浓度小时监测数据按监测时间的先后转换成时间序列，每个站点数据对应一条时间序列；再运用时间序列延迟相关算法对数据进行处理，得到最大延迟相关的时间序列数据，并将其输入 LSTM 模型对目标站点的臭氧浓度进行预测。实验结果表明：本文所提模型的预测结果曲线更为平滑且与真实值更加接近，均方根误差和绝对平均误差均为最小，预测效果相较传统的 LSTM 模型和 SpaceLSTM 模型更好，可以为臭氧的预警预报提供一定的参考。

参考文献

[1] 李汉忠. 第一讲: 臭氧技术的发展[J]. 家用电器, 1998(1): 24-25.

[2] 林璟. 温室效应与臭氧层空洞[J]. 引进与咨询, 2006(6): 108-109.

[3] 张莹, 岳玎利, 江明, 翟宇虹. 珠江三角洲臭氧污染特征与趋势初步分析[J]. 广东化工, 2016, 43(12): 152-153+144.

[4] 严茹莎, 李莉, 安静宇, 等. 上海市夏季臭氧生成与其前体物控制模拟研究[J]. 环境污染与防治, 2016, 38(1): 30-35+40.

[5] 环境保护部. 环境保护部发布 2015 年 12 月重点区域和 74 个城市空气质量状况[J]. 油气田环境保护, 2016(1): 41.

[6] Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A. and Jung, J. (2019) Using a Deep Convolutional Neural Network to Predict 2017 Ozone Concentrations, 24 Hours in Advance. *Neural Networks*, **121**, 396-408. https://doi.org/10.1016/j.neunet.2019.09.033

[7] 王振友, 陈莉娥, 何克尧. 一种改进的 GM(1, 1)预测模型在大气中臭氧含量分析中的应用[J]. 数学的实践与认识, 2007(22): 60-65.

[8] 朱佳, 王振会, 金天力, 郝晓静. 基于小波分解和最小二乘支持向量机的大气臭氧含量时间序列预测[J]. 气候与环境研究, 2010, 15(3): 295-302.

[9] 张春露. 基于 Tensorflow 的 LSTM 在太原空气质量 AQI 指数中的分析与预测[D]: [硕士学位论文]. 太原: 中北大学, 2019.

[10] 林子雨, 江弋, 赖永炫, 林琛. 一种新的时间序列延迟相关性分析算法——三点预测探查法[J]. 计算机研究与发展, 2012, 49(12): 2645-2655.

[11] Sakurai, Y., Papadimitriou, S. and Faloutsos, C. (2005) BRAID: Stream Mining through Group Lag Correlations. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimore, Maryland, USA, 14-16 June 2005, 599-610. https://doi.org/10.1145/1066157.1066226

[12] 杨青, 王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究, 2019, 36(3): 65-77.

[13] 刘礼文, 俞弦. 循环神经网络(RNN)及应用研究[J]. 科技视界, 2019(32): 54-55.

[14] 张萧, 黄晞, 仲伟汉, 等. Sigmoid 函数及其导函数的 FPGA 实现[J]. 福建师范大学学报(自然科学版), 2011(2): 68-71.