Advances in Applied Mathematics 应用数学进展, 2023, 12(5), 2143-2151
Published Online May 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.125218

Hans汉斯

基于LSTM的游客预测

顾永鹏，覃迪波，章　详，崔朝翔，周才英*

江西理工大学理学院，江西　赣州

收稿日期：2023年4月17日；录用日期：2023年5月9日；发布日期：2023年5月16日

摘　要

随着GDP不断的增长，越来越多的人们在节假日愿意将时间用来旅游。景区游客的数量急剧增长和需求的集中，极易造成区域交通的拥堵或景区人挤人、游客之间产生安全事故等现象。严重影响了游客的游玩体验，同时不利于旅游消费和旅游业的可持续发展。为了改善这种现状，本文对各个景区的游客数量按照时间进行预测，通过数据建立模型给游客提供旅游建议。针对上述问题，本文先搜集各个景区旅游数据，在数据量足够大的前提下，使用深度学习中的LSTM模型对数据集进行训练，通过前向传播和反向传播训练模型，从而得到预测结果。然后按照LSTM预测的游客人数，应用在旅游服务平台上，以指导游客错峰出行，提升游客的游玩体验，并有效帮助景区管理人员对商品和人员进行安排。

关键词

LSTM模型，CNN模型，ARIMA模型，SPSS软件

Visitor Forecast Based on LSTM

Yongpeng Gu, Dibo Qin, Xiang Zhang, Zhaoxiang Cui, Caiying Zhou*

School of Science, Jiangxi University of Science and Technology, Ganzhou Jiangxi

Received: Apr. 17th, 2023; accepted: May 9th, 2023; published: May 16th, 2023

Abstract

With the continuous growth of GDP, more and more people are willing to spend their time traveling during holidays. The number of tourists in scenic spots increases rapidly and the demand is concentrated, which is easy to cause regional traffic congestion, crowded scenic spots, safety accidents between tourists and other phenomena. It has seriously affected the tourists' play experience and is not conducive to the sustainable development of tourism consumption and tourism.

*通讯作者。

文章引用: 顾永鹏, 覃迪波, 章详, 崔朝翔, 周才英. 基于 LSTM 的游客预测[J]. 应用数学进展, 2023, 12(5): 2143-2151. DOI: 10.12677/aam.2023.125218

In order to improve this situation, we forecast the number of tourists in each scenic spot according to the time, and build a model through the data to provide tourists with travel suggestions. To solve the appeal problem, we first collect tourism data of each scenic spot, and on the premise of sufficient data, we use the LSTM model in deep learning to train the data set. Through forward propagation and back propagation training model, we can get the predicted results. Then, according to the number of tourists predicted by LSTM, it is applied on the tourism service platform to guide tourists to travel off-peak, improve their playing experience, and effectively help the scenic spot management personnel to arrange commodities and personnel.

Keywords

LSTM Model, CNN Model, ARIMA Model, SPSS Software

1. 引言

随着 GDP 不断的增长，越来越多的人们在节假日愿意将时间用来旅游。旅游业快速发展的同时弊端也逐渐暴露，一些知名景区游客数量在节假日暴涨，导致景区生态环境遭到破坏、更有甚者造成不可估量的人身财产损失。比如最近在国外发生的梨泰院踩踏事件，造成了重大的人员伤亡和受伤事件。因此，控制各景区游客数量对景区发展有重大的作用[1]。

随着我国信息科技的快速发展，在“互联网+”的背景下，人们可以足不出户便可以了解各大景区的详情。旅游的售票方式发生了改变，线上预定售票变成了主流，以及线下售票的数据都上传到景区管理系统中。并且互联网具有实时性，这比用缺乏及时性的传统数据更好，可以防止因重大事件而引起景区游客数量预测的不准确性。比如此次疫情对旅游造成了巨大的冲击，如果用往期数据分析游客数量并不能起到好的预测结果[2]。

2. 前期准备

2.1. 数据可视化与分析

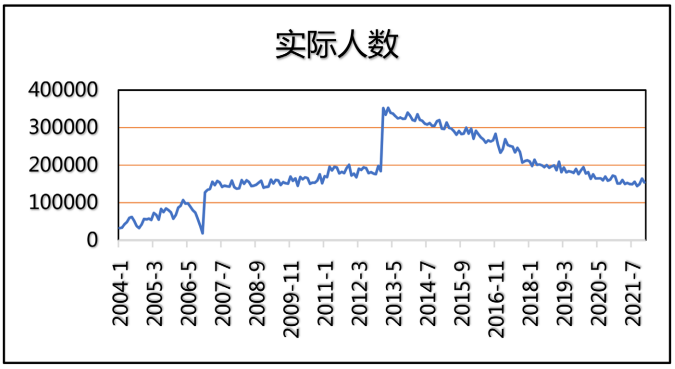

Figure 1. Raw data
图 1. 原始数据

首先本文先将所给的数据进行可视化，得到如下折线图见图 1，其中横坐标为时间，纵坐标为游客人数。

观察图 1 折线图，自 2004 年起到 2021 年本文可以发现游客人数增多了，如果以年为单位划分看数据，发现数据受时间影响，比如 2006 年游客人数持续下滑，2007 到 2013 之间游客人数处于较稳定的波动状态，2014 年到 2021 年之间游客人数持续下降。同时可以看出游客人数在年份之间具有联系，大致呈现出一致的趋势，故采取 LSTM 通过往期数据对未来进行预测。

2.2. LSTM 模型介绍

本文观察到所给的数据是时间序列数据，普通的深度神经网络模型如 FCN、CNN 虽然能够处理输入的时间序列信息，但很难从时间中获取序列间的依赖关系。循环神经网络(Recurrent Neural Network, RNN)是一类专门用于处理序列数据的网络。但容易产生梯度消失、梯度爆炸的问题，于是本文采用优化后的 LSTM 网络模型来处理数据，这是一种改进之后的时间递归神经网络，适用于处理和预测时间间隔较长的长期事件，解决了 RNN 无法处理长距离的依赖的问题。相比与传统 RNN，LSTM 网络增加了遗传门、输入门、输出门与细胞状态，使其可以很好地处理长序列数据。本文采用了 LSTM 模型，其中输入特征维数为 2，隐藏层维度为 4，输出特征维度为 1，在 LSTM 模型由两个 RNN 和一个全连接层组成。

2.3. 与 ARIMA 模型的对比

ARIMA 模型将自回归模型、移动平均模型和差分法结合，得到的差分自回归移动平均模型 ARIMA(p,d,q)，其中 d 是需要对数据进行差分的阶数。模型的基本原理是在将非平稳时间序列转化为平稳时间序列的过程中，将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型，本质上只能捕捉线性关系，不能捕捉非线性关系[3]。因此需要时序数据是稳定的，或者通过差分化之后是稳定的，如果不稳定的数据，将无法捕捉到规律的。考虑到游客人数常常受政策和新闻的影响而波动，故猜测用 ARIMA 无法预测，对此的验证如下：

稳定的数据是没有趋势，没有周期性的；即它的均值，在时间轴上拥有常量的振幅，并且它的方差，在时间轴上是趋于同一个稳定的值的。所以首先对序列绘图，进行 ADF 检验，观察序列是否平稳，使用 SPSS 软件对原始数据进行差分分析，结果见图 2。

由于在 ACF 及偏 ACF 中均存在截尾现象，因此使用 ARIMA 模型的结果较为不理想，需要对参数中的阶数一项进行多次猜测验证才能得到较为标准的 ARIMA 模型，将耗费大量的时间，且根据对原始数据作图观察，发现每月游客人数波动较大，故转而使用 LSTM 模型进行预测。

3. LSTM 模型原理

3.1. LSTM 模型和 RNN 工作原理

LSTM 是 RNN 的特殊类型，RNN(Recurrent Neural Network)是一类用于处理序列数据的神经网络，包含了输入层、隐藏层、输出层。通过激活函数控制输出，层与层之间通过权值连，RNN 不仅在层与层之间建立了权连接，在层之间的神经元之间也建立了权连接，图像展示见图 3。

其中 X 是是输入，h 是隐藏层单元，o 是输出，L 是损失函数，y 是训练集的标签，U，V，W 是权值。

T 时刻：$h^{(t)}=\varnothing\left(Ux^{(t)}+Wh^{(t-1)}+b\right)$

T 时刻输出：$o^{(t)}=Vh^{t}+c$

最终模型输出：$\hat{y}^{(t)}=\sigma\left(o^{(t)}\right)$，图像展示见图 4。

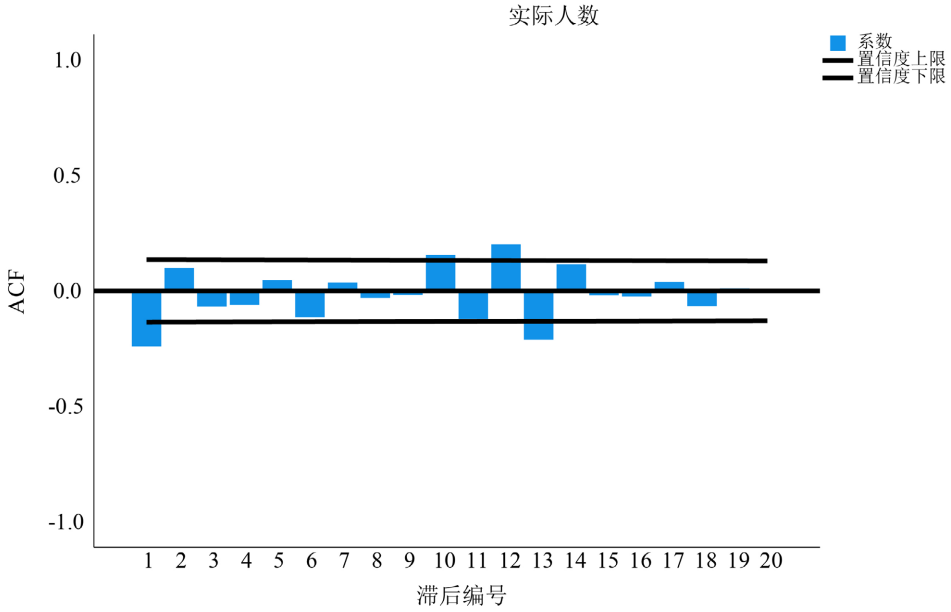

Figure 2. Differential analysis of tourist numbers
图 2. 游客人数差分分析图

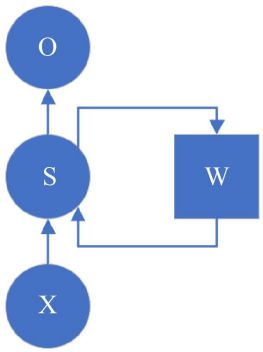

Figure 3. Simple RNN model
图 3. 简单 RNN 模型

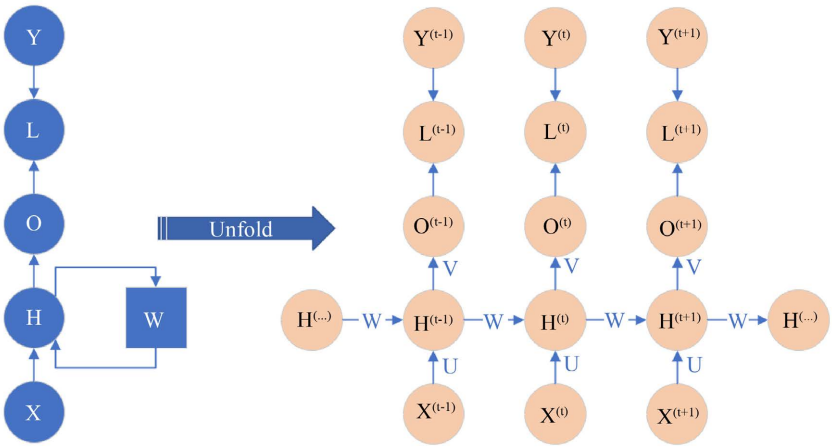

Figure 4. Internal mechanism of RNN
图 4. RNN 内在机理

LSTM 网络通过门控制将短期记忆与长期记忆结合起来，并且一定程度上解决了梯度消失的问题。原始 RNN 的隐藏层只有一个状态，即 h，它对于短期的输入非常敏感。LSTM 再增加一个状态，即 c，让它来保存长期的状态，称为单元状态(cell state) [4]。在 t 时刻，LSTM 的输入有三个：当前时刻网络的输入值 $X_{(t)}$、上一时刻 LSTM 的输出值 $h_{(t-1)}$、以及上一时刻的单元状态 $C_{(t-1)}$；

LSTM 的输出有两个：当前时刻 LSTM 输出值 $h_{(t)}$、和当前时刻的单元状态 $C_{(t)}$，图像展示见图 5。

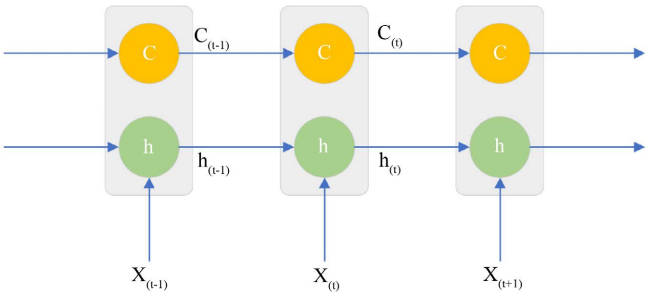

Figure 5. LSTM schematic
图 5. LSTM 原理图

3.2. LTSM 核心思想

LSTM 的关键在于单元状态(cell state)。信息在单元状态(cell state)之间流传保持不变。同时使用三个门来控制长期状态 c [5]。

门(gate)是一层全连接层，输入是一个向量，输出是一个 0 到 1 之间的实数向量。

其中公式为：$g(x)=\sigma(Wx+b)$

方法：用门的输出向量按元素乘以本文需要控制的那个向量。

原理：门的输出是 0 到 1 之间的实数向量，当门输出为 0 时，任何向量与之相乘都会得到 0 向量，这就相当于什么都不能通过；输出为 1 时，任何向量与之相乘都不会有任何改变，这就相当于什么都可以通过。

3.3. LSTM 的计算

遗忘门的计算：它决定了上一时刻的单元状态 $C_{(t-1)}$ 保留到当前时刻 $C_{(t)}$ 的数量，图像展示见图 6。

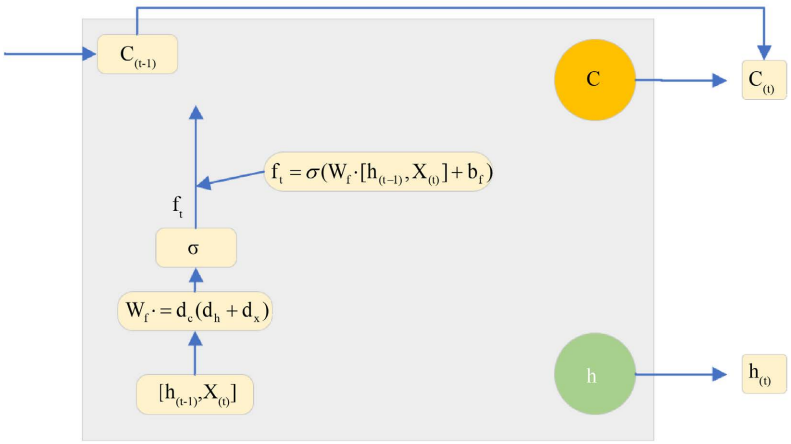

Figure 6. LSTM forgetting gate
图 6. LSTM 遗忘门

其中W_f是遗忘门的权重矩阵，$\left[h_{(t-1)}, X_{(t-1)}\right]$将两个向量连接成一个更长的向量，$b_f$是遗忘门的偏置项，$\sigma$是sigmoid函数。

输入门的计算：输入门决定了当前时刻网络的输入$X_{(t)}$保存到单元状态$C_{(t)}$的数量，可以避免无关的内容进入记忆，图像展示见图 7。

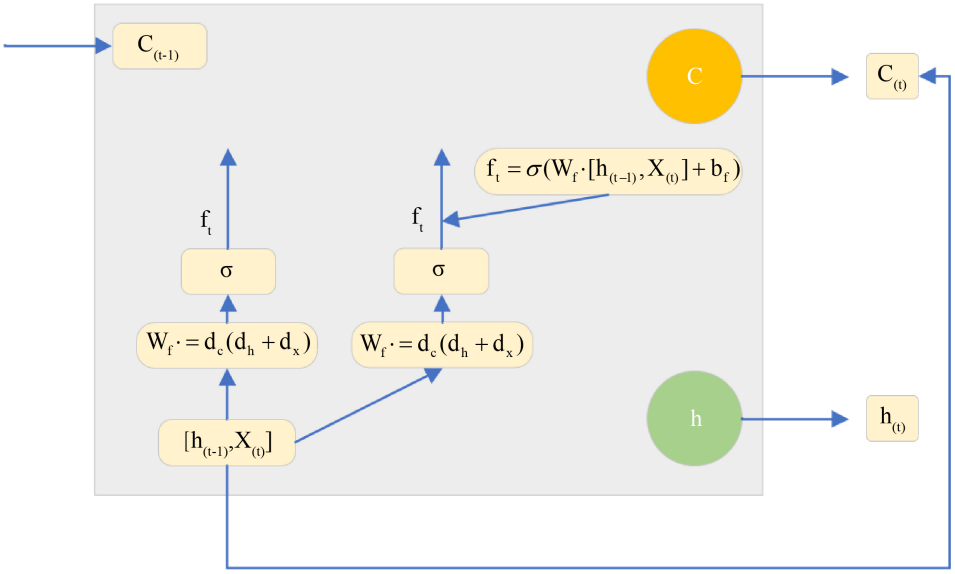

Figure 7. LSTM input gate
图 7. LSTM 输入门

根据上一次的输出和本次输入来计算当前输入的单元状态，图像展示见图 8。

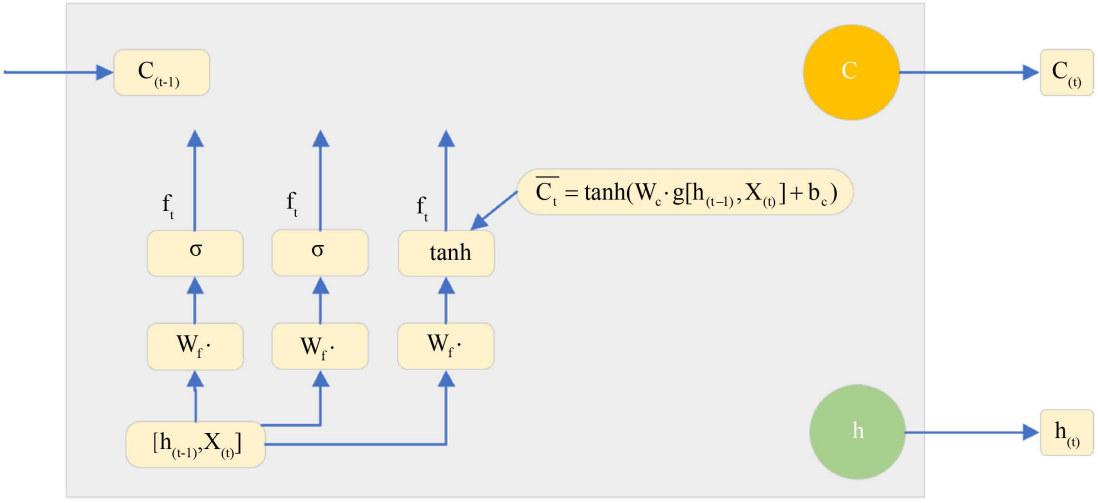

Figure 8. LSTM output gate
图 8. LSTM 输出门

当前时刻的单元状态$C_{(t)}$的计算：由上一次的单元状态$C_{(t-1)}$按元素乘以遗忘门f_t，再用当前输入的单元状态$C_{(t)}$按元素乘以输入门i_t，再将两个积加和：这样，就可以把当前的记忆$C_{(t)}$和长期的记忆$C_{(t-1)}$

组合在一起，形成了新的单元状态$C_{(t)}$。由于遗忘门的控制，它可以保存很久很久之前的信息，由于输入门的控制，它又可以避免当前无关紧要的内容进入记忆，图像展示见图 9。

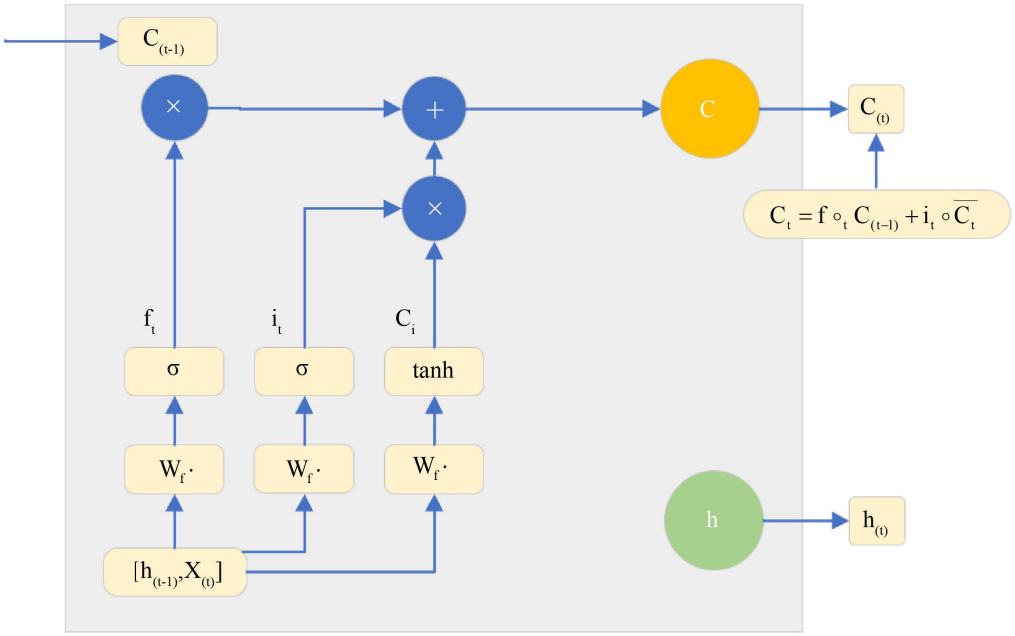

Figure 9. Calculation of the LSTM input gate
图 9. LSTM 输入门的计算

输出门的计算：

输出门用来控制单元状态$C_{(t)}$输出到 LSTM 的当前输出值$h_{(t)}$的数量，控制了长期记忆对当前输出的影响，图像展示见图 10。

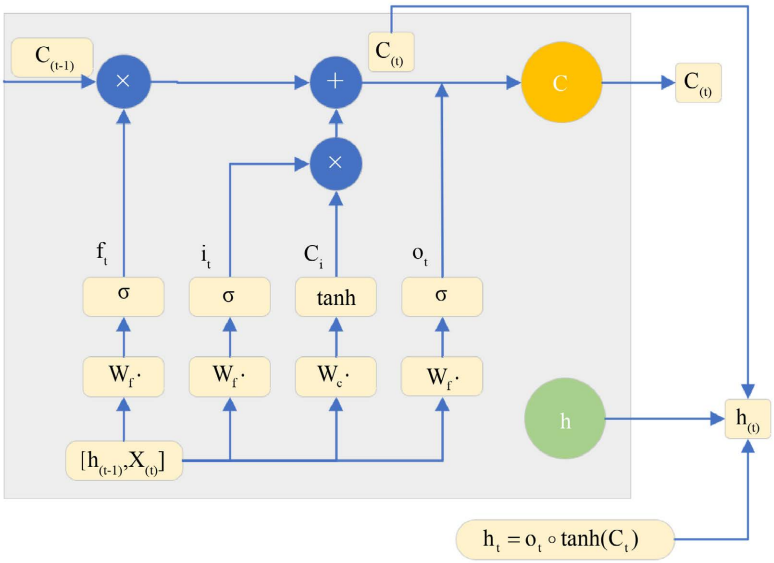

Figure 10. Calculation of the LSTM output gate
图 10. LSTM 输出门的计算

4. 模型的结果

基于游客人数数据量较大的情况，通过 LSTM 模型发现拟合的数据和实际的数据基本吻合，并且预测了 2022 年游客人数的情况。每次都会以训练集：测试集 = 7:3 的比例来训练数据，从而确保模型的泛化能力[6]。本文最终得出了游客人数的预测数值，其中红色为预测数据，蓝色为原始数据，最终结果图片见图 11，具体预测人数见表 1。

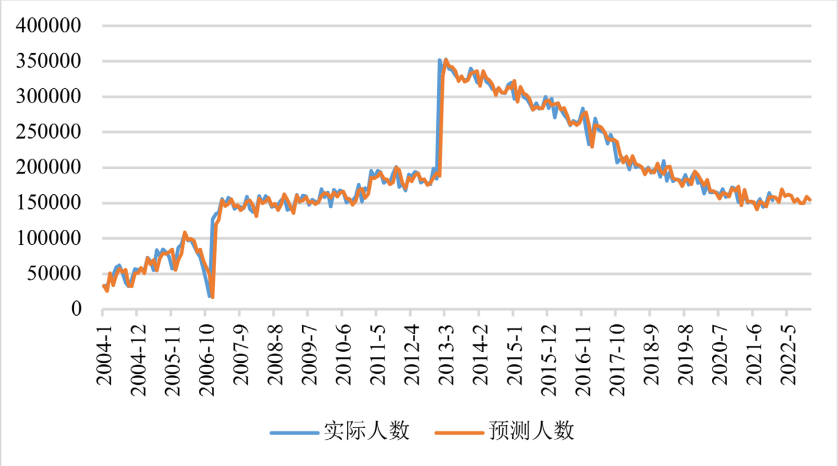

Figure 11. Forecast results of tourist arrivals
图 11. 游客人数预测结果

通过上述 LSTM 模型，可以看出使用 LSTM 模型对游客人数的预测较为理想，且对于大多数波动情况均能及时的反应，因此对于预测每月的游客人数的可行性较高，故可以使用预测的游客人数进行对景区和游客等给予相关的参考数据，从而帮助其进行决策，如给游客一些旅行时间的建议或用于景区管理人员合理调配资源，景区及周边商家的价格调整等。

Table 1. Tourist population projections
表 1. 游客人数预测数据

年月	游客人数/万人
2022-1	157,952
2022-2	151,307
2022-3	169,420
2022-4	159,977
2022-5	162,078
2022-6	160,539
2022-7	151,582
2022-8	155,893
2022-9	149,756
2022-10	149,916
2022-11	159,205

5. 模型的缺点与改进

模型的缺点：

1) Time consuming：由于训练 lstm 模型需要根据结果不断通过前向传播和反向传播修改系数以达到较好的预测效果，这需要花费大量时间[7]。

2) 在使用 LSTM 模型进行预测时，以月为单位预测，使得结果不够精确，存在一定的误差。

模型的优点：

1) lstm 能很好地处理时间序列，并且 lstm 相较于 rnn 不会出现梯度消失、梯度爆炸等问题。

2) 以一个月进行预测便于景区进行调控，同时避免因短期重大事件而导致预测数据和实际数据误差过大的情况。

6. 结束语

本文直接利用原始数据，预测出未来一年的游客人数变化情况较为理想，拟合曲线平滑，但少数时候的突变导致的噪声点有极大的可能造成剧烈的影响，且以月份为单位进行预测数据较为模糊，故模型并不能得到完美的预测值，但对规避不必要且致命的风险时却有着极好的效果。

参考文献

[1] 时萍萍, 胡姚刚, 孟继东. 基于互联网旅游数据的游客量预测模型研究现状与展望[J]. 资源开发与市场, 2022, 38(8): 921-929.

[2] 邓雨菲. ARIMA-ATT-LSTM 在旅游客流量预测中的应用研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2022. https://doi.org/10.26991/d.cnki.gdllu.2022.000310

[3] 丁锐, 李伟, 王若舟. 基于 SARIMA 和 LSTM 组合预测模型包量[J]. 计算机与数字工程, 2020(2): 304-307.

[4] 张晨阳, 韦增欣, 郜星军. 基于 LSTM 模型的数学机理分析实证研究包量[J]. 中国管理信息化, 2019(15): 93-97.

[5] 崔国超. 神经网络模型特性研究包量[J]. 无线互联科技, 2012(3): 105.

[6] 陈波杰, 蔡乐才, 刘星, 成奎. 一种优化 LSTM 神经网络模型的预测方法包量[J]. 四川轻化工大学学报: 自然科学版, 2022(5): 78-86.

[7] 周金荣, 黄道, 蒋慰孙. 一种新型神经网络模型的研究[J]. 信息与控制, 1994(1): 22-27.