主题：Time series analysis based on deep learning

张靖哲 21009200835

实验目标：

使用lstm的深度学习算法对时间序列进行分析，并使用测试集尝试复现股票价格变化趋势

引言

LSTM（Long Short-Term Memory）是一种常用的循环神经网络（RNN）模型，用于处理序列数据，具有记忆长短期的能力。在时间序列预测中，LSTM既可以多元预测机制又可以作为单元预测机制使用。

作为多元预测机制，LSTM可以处理多个相关变量的历史数据，从而可以预测这些变量的未来值。具体地，我们可以将多个变量的历史数据作为LSTM的输入，将多个变量的未来值作为LSTM的输出。在训练过程中，我们可以使用误差反向传播算法来更新LSTM的参数，从而优化模型的预测性能。

作为单元预测机制，LSTM可以预测单一变量的未来值，例如股票价格、销售量等。在单元时间序列预测中，我们需要对历史数据进行分析，确定趋势、季节性和周期性等因素，并使用这些因素来预测未来的值。LSTM可以通过学习历史数据中的模式和规律，来预测未来的值。

LSTM作为多元预测机制和单元预测机制的优点是可以处理序列数据中的长期依赖关系，从而可以捕捉到数据中的复杂模式和规律。它可以自适应地学习和调整模型参数，从而提高模型的预测性能和泛化能力。

总的来说，LSTM作为多元预测机制和单元预测机制的应用广泛，可以用于预测股票价格、气象数据、交通流量等多个领域的数据。

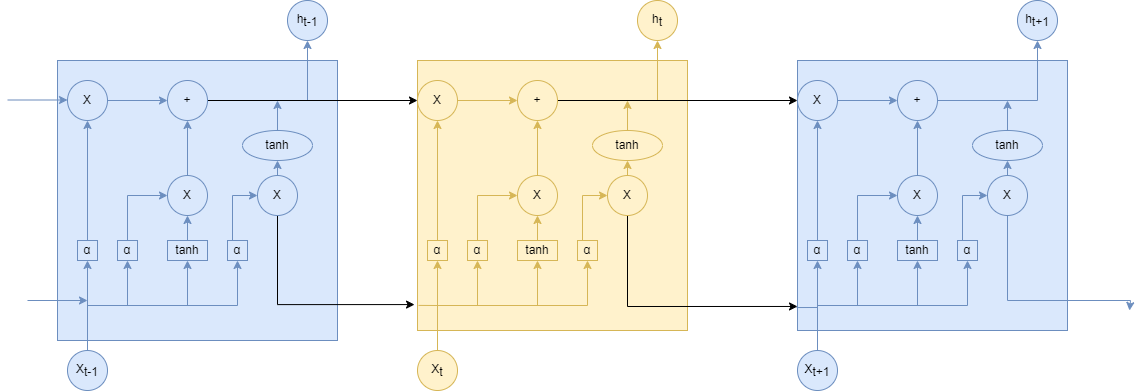
为什么需要LSTM

LSTM从被设计之初就被用于解决一般递归神经网络中普遍存在的长期依赖问题，使用LSTM可以有效的传递和表达长时间序列中的信息并且不会导致长时间前的有用信息被忽略（遗忘）。与此同时，LSTM还可以解决RNN中的梯度消失/爆炸问题。

LSTM机制  
LSTM（长短期记忆，Long Short-Term Memory）是一种用于处理序列数据的深度学习模型，属于循环神经网络（RNN）的一种变体,其使用一种类似于搭桥术结构的RNN单元。相对于普通的RNN，LSTM引入了门控机制，能够更有效地处理长期依赖和短期记忆问题，是RNN网络中最常使用的Cell之一。

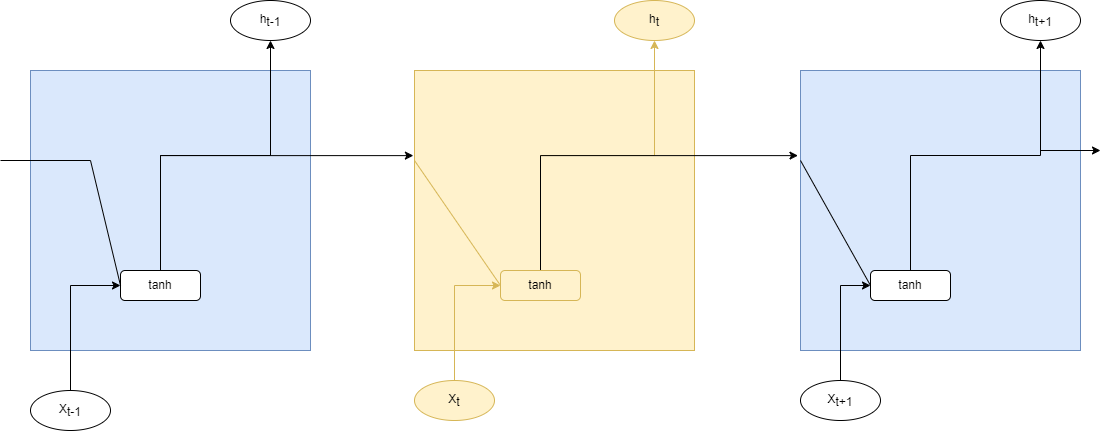
1、了解LSTM的结构

LSTM通过刻意的设计来实现学习序列关系的同时，又能够避免长期依赖的问题。它的结构示意图如下所示。



在LSTM的结构示意图中，每一条黑线传输着一整个向量，从一个节点的输出到其他节点的输入。其中“+”号代表着运算操作(如矢量的和)，而矩形代表着学习到的神经网络层。汇合在一起的线表示向量的连接，分叉的线表示内容被复制，然后分发到不同的位置。

如果上面的LSTM结构图你看着很难理解,但是其实LSTM的本质就是一个带有tanh激活函数的简单RNN，如下图所示



LSTM这种结构的原理是引入一个称为细胞状态的连接。这个状态细胞用来存放想要的记忆的东西(对应简单LSTM结构中的h，只不过这里面不再只保存上一次状态了，而是通过网络学习存放那些有用的状态)，同时在加入三个门,分别是：

     忘记门:决定什么时候将以前的状态忘记。

     输入门：决定什么时候将新的状态加进来。

     输出门：决定什么时候需要把状态和输入放在一起输出。

从字面上可以看出，由于三个门的操作，LSTM在状态的更新和状态是否要作为输入，全部交给了神经网络的训练机制来选择。

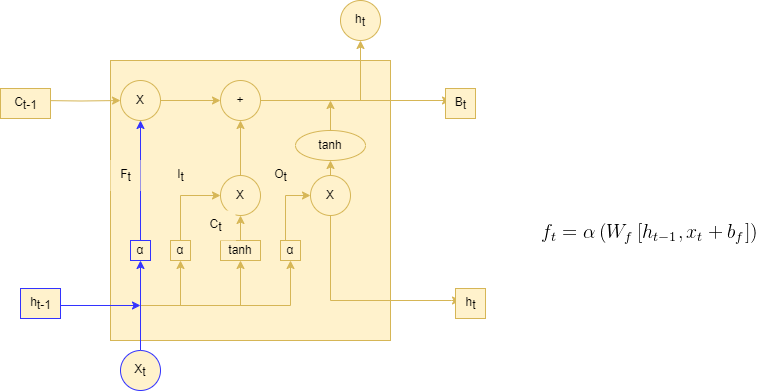
下面分别来介绍一下三个门的结构和作用。

（1）忘记门

下图所示为忘记门的操作，忘记门决定模型会从细胞状态中丢弃什么信息。

忘记门会读取前一序列模型的输出和当前模型的输入来控制细胞状态中的每个数是否保留。

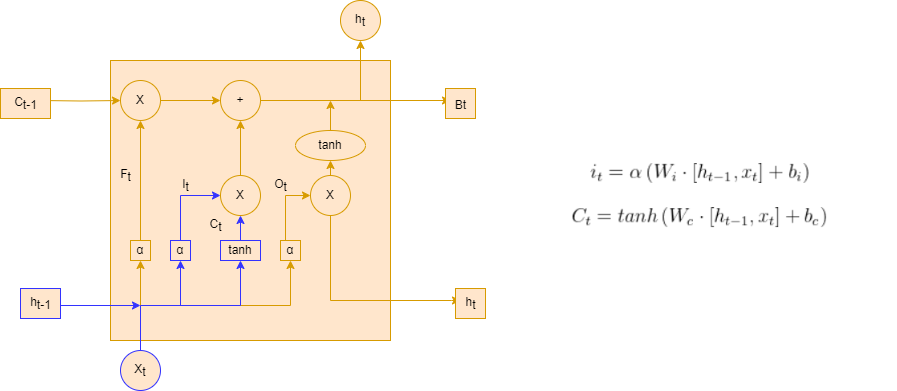
例如:在一个语言模型的例子中，假设细胞状态会包含当前主语的性别，于是根据这个状态便可以选择正确的代词。当我们看到新的主语时，应该把新的主语在记忆中更新。忘记们的功能就是先去记忆中找到一千那个旧的主语(并没有真正执行忘记的操作，只是找到而已。



在上图的LSTM的忘记门中，代表忘记门的输出， α代表激活函数，代表忘记门的权重，代表当前模型的输入，代表前一个序列模型的输出，代表忘记门的偏置。

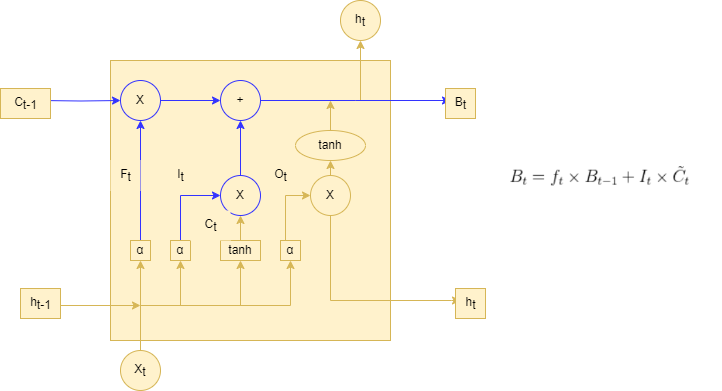
（2）输入门

输入门可以分为两部分功能，一部分是找到那些需要更新的细胞状态。另一部分是把需要更新的信息更新到细胞状态里



在上面输入门的结构中，代表要更新的细胞状态，α代表激活函数，代表当前模型的输入，代表前一个序列模型的输出，代表计算的权重，代表计算的偏置，代表使用tanh所创建的新细胞状态，代表计算的权重，代表计算的偏置。

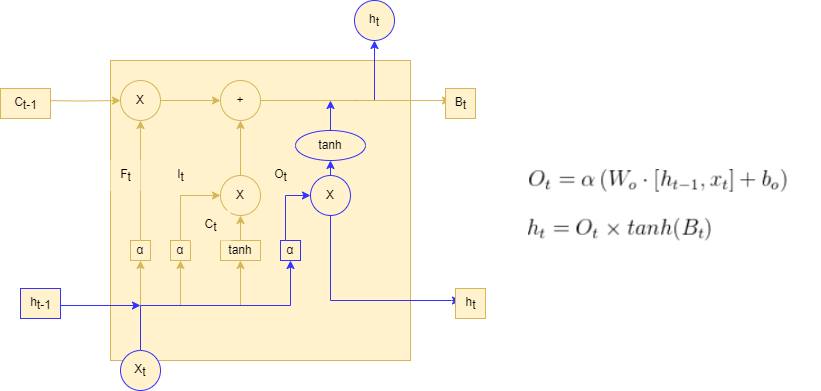
忘记门找到了需要忘掉的信息后，在将它与旧状态相乘，丢弃确定需要丢弃的信息。(如果需要丢弃对应位置权重设置为0)，然后，将结果加上 \* 使细胞状态获得新的信息。这样就完成了细胞状态的更新，如下图输入门的更新图所示。



上图LSTM输入门的更新图中，代表忘记门的输出结果， 代表忘记门的输出结果，代表前一个序列模型的细胞状态，代表要更新的细胞状态，代表使用tanh所创建的新细胞状态。

（3）输出门

如下图LSTM的输出门结构图所示，在输出门中，通过一个激活函数层(实际使用的是Sigmoid激活函数)来确定哪个部分的信息将输出，接着把细胞状态通过tanh进行处理(得到一个在-1~1的值)，并将它和Sigmoid门的输出相乘，得出最终想要输出的那个部分，例如，在语言模型中，假设已经输入了一个代词，便会计算出需要输出一个与该代词相关的信息(词向量)



在LSTM的输出门结构图中，代表要输出的信息，α代表激活函数，代表计算 的权重，代表计算的偏置，代表更新后的细胞状态，代表当前序列模型的输出结果。

实验过程：

选用的数据集是Kaggle 上的 Stock Price Prediction 数据集。

选取日期和每日最高股价作为分析目标：

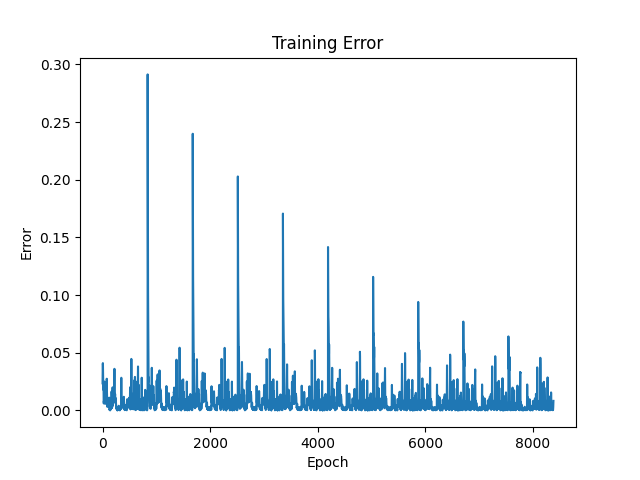
dates = pd.to\_datetime(true\_data['Date'])  # 提取日期信息

target = 'High'

并把股价做归一处理。

设置Train标志为True，进行模型训练。使用均方误差（MSE）作为损失函数，Adam优化器进行参数优化。

训练过程中，输出每个epoch的损失值，并保存训练好的模型。

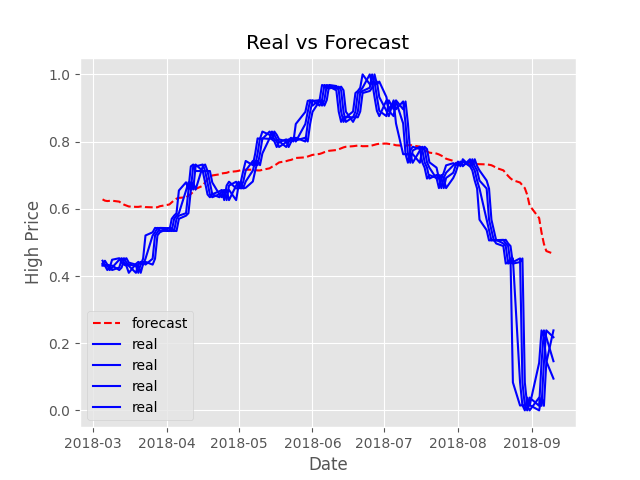


上图是LSTM训练后的误差损失图。

然后设置Train标志为False，加载已训练好的模型，并在测试集上进行预测。

计算预测结果与真实值之间的平均绝对误差（MAE）。

输出预测结果，并绘制实际值和预测值的对比图。



可见根据较少的测试集模型能大致预测出股票的趋势，但在拐点有一定的滞后性。

实验总结：

LSTM是一种适用于处理长时依赖性的循环神经网络（RNN）变体。对于时间序列数据而言，过去的信息可能对当前和未来的预测具有重要影响。LSTM能够捕捉和利用这种长时依赖性，有助于更好地建模时间序列的动态变化。

LSTM模型对数据中的噪声和不规则模式具有一定的鲁棒性，能够从中学到趋势和模式。

LSTM的性能很大程度上取决于超参数的选择，包括学习率、网络层数、隐藏单元数量等。因此，在应用中，需要仔细调参以获得最佳性能。

总体而言，LSTM在处理时间序列数据方面表现出色，特别是在需要考虑长期依赖关系和复杂模式的情况下。