课程项目设计报告

项目名称：**Ubiquant市场预测——**基于未来市场数据预测



报告人： 黄俊卿 陆奕丞 陈庭柱

报告时间： 2023年09月01日

1 研究背景和项目目标

1.1选题依据

随着经济全球化和信息技术的发展，股票市场已成为全球经济的重要组成部分。对股票市场的准确预测和分析对于投资者、基金经理和政策制定者等都具有重要意义。然而，股票市场具有非线性特征，受多种因素影响，包括社会、经济、舆论、媒体、交易行为、政治等，这使得预测和分析变得十分复杂和困难。

在这种背景下，机器学习技术开始被应用于股票市场预测和分析。机器学习可以通过对大量历史数据的学习，发现数据中的模式和规律，并根据这些模式和规律对未来进行预测。机器学习的应用可以帮助投资者更加准确地预测股票市场的走势和个股股价的变化，从而做出更加明智的投资决策，减少投资风险，提高投资收益。

1.2业界现状介绍

在过去的几年里，人工智能技术在许多领域都取得了显著的突破，其中包括股票预测。许多研究人员和公司已经利用深度学习、机器学习和其他先进技术来预测股票收益。以下是一些最新的成果：

1. 基于深度学习的股票预测模型：这些模型利用深度神经网络来预测股票收益。一种名为“深度信念网络”（Deep Belief Networks）的算法在处理股票数据时表现出良好的性能。它能够识别出隐藏在时间序列数据中的模式，并预测未来的股票价格。
2. 基于社交媒体的股票预测：一些研究者发现，社交媒体上的情感和情绪可以提供有关股票市场的有价值信息。通过分析社交媒体上的情感和关键词，一些模型能够预测股票市场的波动，从而帮助投资者做出更明智的投资决策。
3. 利用区块链技术的股票预测：区块链技术为股票预测提供了新的可能性。一些公司正在探索使用区块链技术来创建一个更加透明和高效的股票市场。通过使用智能合约和去中心化的数据源，这些平台可以提供更加准确的股票预测。
4. 基于机器学习的量化交易策略：许多金融机构已经开始使用机器学习技术来开发量化交易策略。这些策略利用历史数据和算法模型来预测股票价格的变动，从而帮助投资者获得更高的回报。

总之，人工智能在股票预测方面已经取得了显著的进展。然而，这些技术仍然存在一定的局限性和风险。投资者应该仔细评估这些方法的有效性，并结合其他指标和策略来做出更加明智的投资决策。

1.3 本项目的目标

小组成员根据课程提供的数据，采用包括但不限于传统机器学习与深度神经网络等方法建立数学模型，并完成以下任务：

(1) 数据分析：对数据进行清理，并开展探索性数据分析，包括但不限于采用图形、表格、动画等展示有用的信息，并给出初步的分析

(2) 抽象数学模型：根据分析结果抽象数学模型，并采用包括但不限于传统机器学习与深度神经网络等一种/多种方法建立初步模型，对未来1天的A股收益进行预测

(3) 进行回测：使用 qlib 或者 backtrader 对相关方法进行回测，并利用得到的数据验证模型合理性

(4) 实验与改进：分析预测结果并根据结果对模型做出相应改进

2 项目总体设计

2.1 项目模块

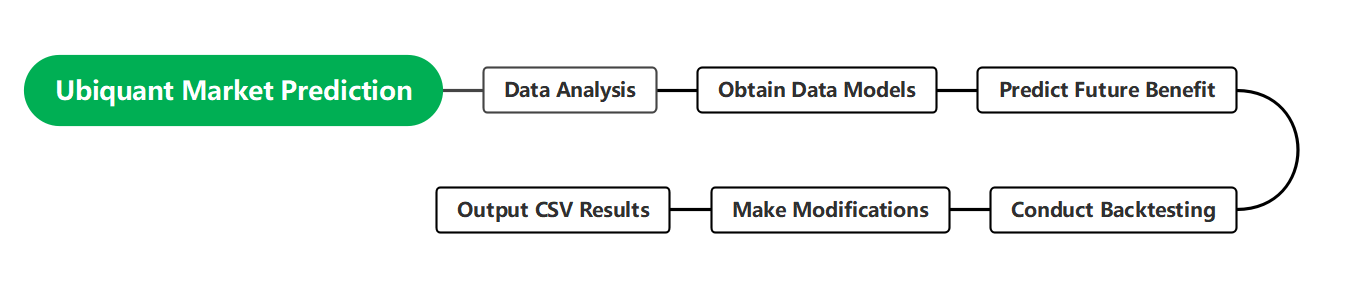
小组成员根据课程提供的数据，采用包括但不限于传统机器学习与深度神经网络等方法建立数学模型，并完成以下任务：

(1)对数据进行清理，并开展探索性数据分析，并给出初步的分析；

(2)根据分析结果抽象数学模型，并采用包括但不限于传统机器学习与深度神经网络等一种/多种方法建立初步模型，根据提供的股票收益数据，由前一日收益预测后一日；

(3)分析预测结果并根据结果对模型做出改进；

(4)以Backtrader量化投资模型为基础，设计一个Strategy，使得在特定的时间段内进行自动买入卖出交易，并给出最后回测时的总收益。



3 项目关键技术

3.1 项目工具

在建模分析模块中，使用python作为语言，通过调用包括机器学习与神经网络等多种模型实现对数据的预测。

在回测分析中，以Backtrader量化投资模型为基础，设计一个Strategy运行，预测特定时间段之后，交易的总收益。

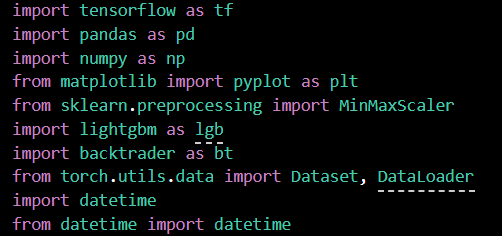
3.2 项目引用库、模型

本项目中，采用的库有：

①numpy；②pandas；③sklearn；④matplotlib；⑤tensorflow；⑥ backtrader ；⑦ datetime；⑧prophet⑨ lightgbm；⑩ torch

本项目后一日股票收益的预测模型为：

LSTM长短时记忆神经网络架构、LGBM基于梯度增强决策树算法的机器学习算法

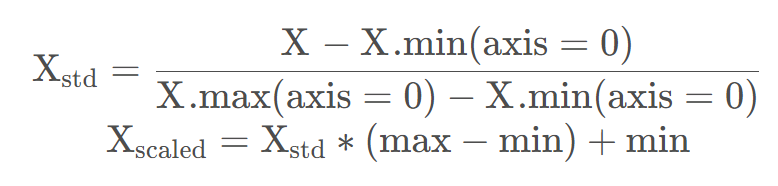


库引用

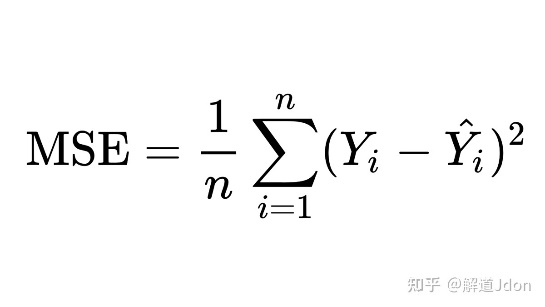
3.3 项目引用算法

本项目中，采用的算法有：

1. minmaxscaler：数据归一化；



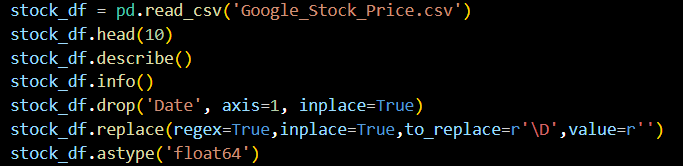
1. mse:均方误差：



4 项目实现

4.1收益预测模型

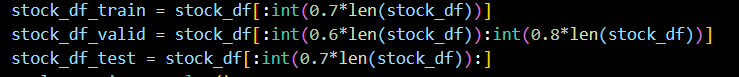
4.1.1 预处理输入数据



预处理合并数据

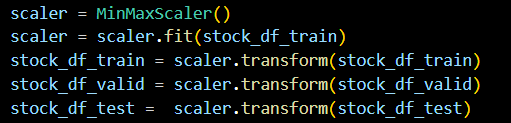
从两张表中分别读取date、high、low等数据。

4.1.2 分离train、valid、test集



分集操作

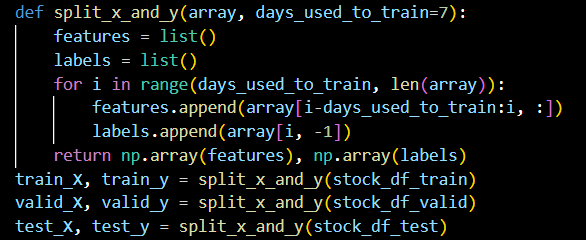
4.1.3 数据归一化



归一化数据

使用MinMaxScaler函数对数据进行归一化处理。

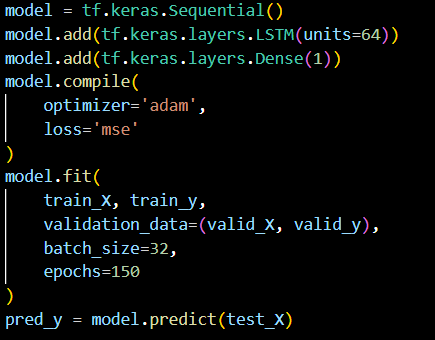
4.1.4 分离label-feature



分离XY

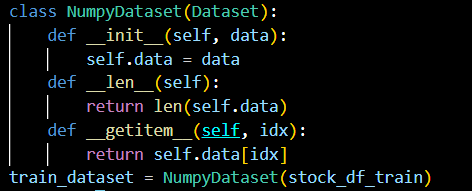
4.1.5 模型的建立与编译

4.1.5.1 建立LSTM模型函数并调用



建立并调试LSTM模型函数

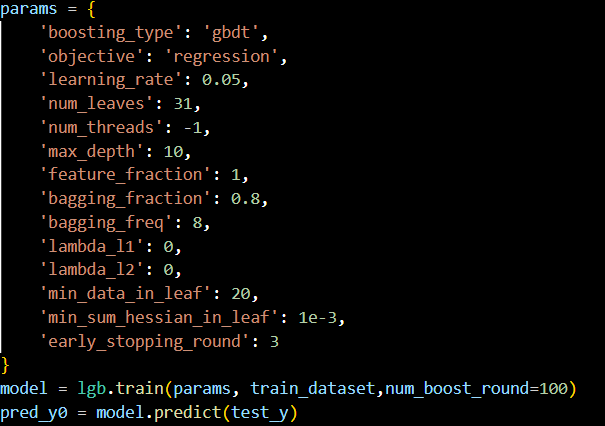
4.1.5.2 建立LGBM模型函数并调用



转换对象格式

Lgb.train() 只接受 Dataset 对象，但stock\_df\_train的类型是 ndarray

所以我们需要把他转换为一个 Dataset 类型对象。



建立并调试LGBM模型函数

boosting\_type: 这个参数指定了boosting类型。  
objective: 这个参数指定了学习任务的目标。  
learning\_rate: 这个参数是学习率，它决定了每一步的权重。  
num\_leaves: 这个参数指定了决策树的叶子节点数。num\_threads: 这个参数指定了用于并行计算的线程数。max\_depth: 这个参数指定了决策树的最大深度。feature\_fraction: 这个参数是一个在0和1之间的数值，表示随机选择的特征的比例。  
bagging\_fraction: 这个参数是一个在0和1之间的数值，表示用于bagging的样本的比例。

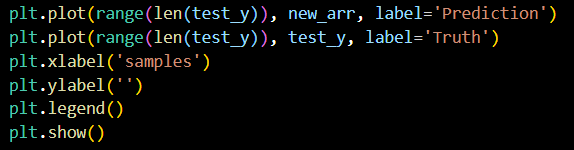
lambda\_l1: 这个参数是一个正数，表示L1正则化项的系数。  
lambda\_l2: 这个参数是一个正数，表示L2正则化项的系数。  
min\_data\_in\_leaf: 这个参数指定了叶子节点所需的最小样本数。

min\_sum\_hessian\_in\_leaf: 这个参数指定了叶子节点所需的最小二阶导数的和。

early\_stopping\_round: 这个参数表示在验证误差没有提升时提前停止训练的轮数。

params：之前定义的参数字典。  
train\_x：训练集的特征矩阵。  
train\_y：训练集的目标变量。  
num\_boost\_round，它指定了梯度提升的迭代次数。

4.1.6 绘制图像



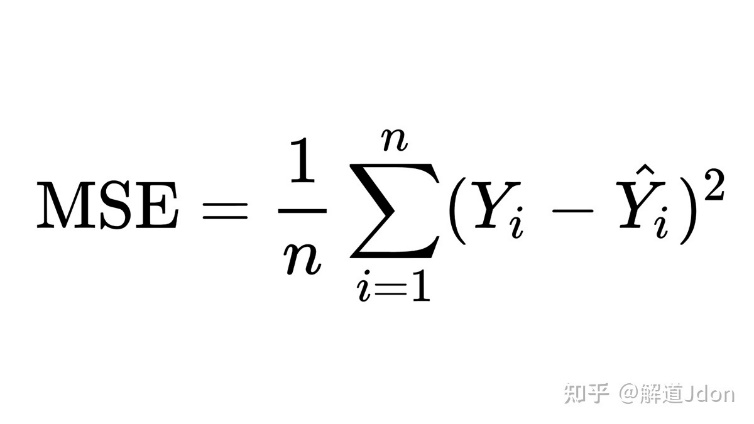
完成预测后，绘制图形

使用Matplotlib库来绘制一条预测曲线和一条真实值曲线，用于展示模型的预测效果。通过这个图形，我们可以直观地看到模型对测试数据的预测效果，以及预测值与真实值之间的差距。

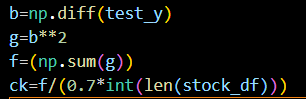
4.1.7 MSE误差计算

由于股票预测具有时效性，所以我们采取“后一日预测收益=前一日实际收益”作为参考误差值。

本项目采取均方误差进行误差计算

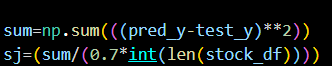


均方误差法算法



参考误差计算

在这段代码中，test\_y是一个包含实际值的数组，np.diff(test\_y)计算了相邻元素之间的差分，即变化量。将这些变化量平方后，得到g数组。  
然后，使用np.sum(g)将g数组中的所有元素求和，得到总平方误差。  
最后，将总平方误差除以测试集总长度，得到参考MSE



实际误差计算

首先，计算了预测值（pred\_y）和测试值（test\_y）之间的平方差的累加和。  
然后，计算了平均平方误差。

|  |  |  |
| --- | --- | --- |
| **采用模型** | **参考MSE** | **实际MSE** |
| LSTM | 0.006288766239623878 | 0.013882809353682779 |
| LSTM/LGBM | 0.005504006490307155 |

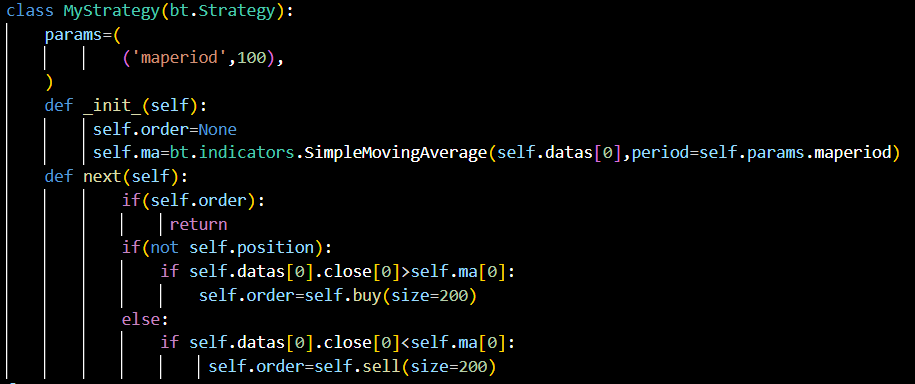
MSE计算结果

可知，我们采取的LSTM/LGBM模型有一定合理性。

4.2 回测收益预测

Backtrader是一个基于Python语言的进行自动化回溯测试的平台。它可用于回测交易策略和模拟实际交易情况。这个工具可以添加自定义的指标和交易策略，提高对交易系统回测的效率。Backtrader可以导入自己的行情数据文件，也可以添加自定义的指标。测试结束后能显示指标和行情图表，而且可以对指标的不同参数设置进行批量测试。

* + 1. 回测战略设置

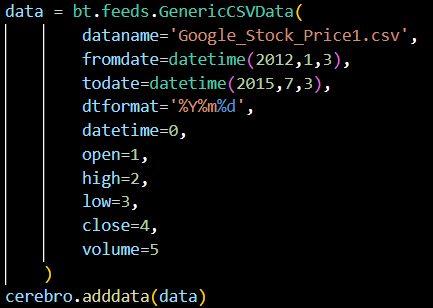


战略设置

策略的工作流程如下：

1. 在初始化阶段 (\_init\_)，计算长期移动平均线（MA）。
2. 在每个交易日，检查是否有未执行的订单 (self.order)。如果有，则返回，不进行任何操作。
3. 如果当前没有持仓（self.position 为 False），并且当前收盘价高于短期移动平均线，则买入（self.buy(size=200)）。这里的 size=200 表示买入200股。
4. 如果当前有持仓（self.position 为 True），并且当前收盘价低于短期移动平均线，则卖出（self.sell(size=200)）。

4.2.2 回测数据导入



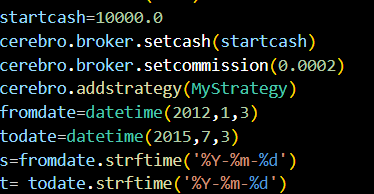
数据导入

这段代码创建了一个GenericCSVData对象，该对象从指定的CSV文件中加载数据。以下是各个参数的含义：

* dataname：CSV文件的名称。
* fromdate和todate：想要加载的数据的时间范围。这里设定为从2012年1月3日到2015年7月3日。
* dtformat：日期格式。这里使用的是年-月-日的格式（例如"20120103"）。
* datetime：在CSV文件中，日期字段的索引。
* open, high, low, close, volume：分别对应股票的开盘价、最高价、最低价、收盘价和交易量。这些是字段在CSV文件中的索引。

这段代码创建的GenericCSVData对象可以传递给Backtrader的策略进行进一步的处理和分析。

4.2.3 回测交易配置



回测交易配置

使用Backtrader进行回测交易的配置。

cerebro.adddata(data)：这行代码向Cerebro添加了一个数据源。

startcash=10000.00：这行代码设置了初始现金为10000.00。

cerebro.broker.setcash(startcash)：这行代码设置了Cerebro的经纪人的初始现金为刚才定义的startcash。

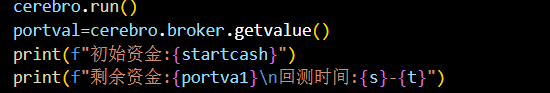
cerebro.broker.setcommission(0.0002)：这行代码设置了Cerebro的经纪人的佣金率为0.0002。

cerebro.addstrategy(MyStrategy)：这行代码向Cerebro添加了一个策略对象。

fromdate=datetime(2012,1,3)：这行代码定义了回测的起始日期，即从2012年1月3日开始回测。

todate=datetime(2015,7,3)：这行代码定义了回测的结束日期，即回测到2015年7月3日。

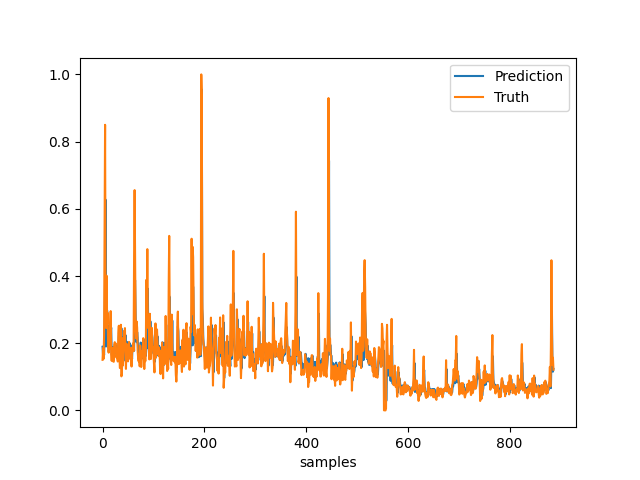
4.2.4回测运转



这段代码首先获取了开始日期（fromdate）和结束日期（todate），并将它们转换为字符串格式，格式为YYYY-MM-DD  
然后，cerebro.run()是开始执行回测的过程。  
在回测结束后，cerebro.broker.getvalue()会返回当前经纪人的账户余额。  
最后，代码打印出初始资金、剩余资金和回测的时间范围。

5 项目测试

5.1股票收益模型预测结果

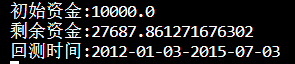


测试集结果

5.2 回测结果

|  |  |  |  |
| --- | --- | --- | --- |
| **Fromdate** | **Todate** | **Startcash** | Backtesting Results |
| 2012-01-03 | 2015-07-03 | 10000.0 | 27687.861271676302 |
| 2012-01-03 | 2014-07-03 | 10000.0 | 21907.514450867053 |
| 2012-01-03 | 2015-07-03 | 15000.0 | 41531.791907514445 |
| 2013-01-03 | 2015-07-03 | 10000.0 | 19745.735472942925 |
| 2013-01-03 | 2014-07-03 | 20000.0 | 32254.335260115608 |

回测结果



回测示例

6 项目管理

6.1 团队人员组成

项目组长：陈庭柱

项目成员：陈庭柱、黄俊卿、陆奕丞

6.2 任务分工

陈庭柱：组长，负责LGBM模型分析、数据选取、修改完成了光伏的建模预测、进行答辩。

黄俊卿：负责建模分析、使用backtrader 对模型进行运行回测并提出优化、报告与PPT的撰写。

陆奕丞：负责建模分析、搭建了A股收益预测模型的构架并进行后续模型的拟合与修改

7. 总结与反思

7.1 项目优点

1.项目中，我们在进行建模时能够自主探索，对于同一个问题，我们能够通过建立多种模型进行预测，最终挑选出最佳的、能够运行的模型并得出结果，使预测数据结果能够尽可能贴近正确。

2.各个模块内，小组成员通力合作，能够有效推进项目进度，不断完善代码，使得最终的预测结果更为合理准确。

7.2 项目不足

1.在项目中，模块之间对接不够清晰，导致模块整合时，出现较大失误，最终通过小组内部会议解决矛盾，能够继续推进。同时，分工不是非常合理导致部分小组成员推进速度过慢。

2.在项目中，我们对Backtrader的使用很不熟练，战略编写并不是十分合理，还需要对他进行进一步的研究。

3.我们对模型预测值的拟合度还不是很合理，在以后的项目中，我们需要训练一个可以找出最合理拟合度的模型，防止拟合程度不够或者过拟合出现。

7.3 小结

在本次项目中，我们面临的主要挑战是如何在有限的时间内训练出一个高效准确的模型。尽管我们明白，为了得到最好的结果，需要充分的训练时间和合适的参数配置，如批次、轮次、算法选择和训练集大小等，但因为时间的限制，我们无法找出最完美的设置。

我们的团队以高效率和决心应对这一挑战。我们快速地收集和处理数据，并立即开始训练模型。尽管时间紧迫，但我们依然尽力探索各种可能的参数组合。我们尝试了不同的批次大小、轮次数量，以及不同的算法，并使用了各种可用的训练集大小。在我们的研究中，我们首先选择使用LSTM模型，然而，我们发现其效果并不理想。主要问题表现在对高点的拟合效果较差，这在一定程度上影响了我们的模型表现,同时，这一部分误差也使得我们的MSE计算误差较大。经过深入探讨和反复实验，我们决定采用一种新的方法，即将LSTM和LGBM（LightGBM）相结合，共同输出我们的预测结果。基于这样的模型架构，我们可以利用LSTM在处理序列数据上的优势，同时利用LGBM在处理非线性关系和高频数据上的优势。通过调整模型参数和优化模型结构，我们成功地将这两种模型进行了有效的融合。在我们的测试中，这种新的模型结构表现出了更好的性能，对高点的拟合效果也有了明显的改善。总的来说，我们的研究显示，将LSTM和LGBM结合使用可以有效地提高模型的性能，期待在未来的工作中，我们将进一步优化这个模型，并对其进行更深入的评估和应用。

然而，由于时间的限制，我们无法对每一个可能的参数组合进行详细的试验和评估。我们尚未找出最优的批次大小、轮次数量，或是最佳的算法选择和训练集大小。但我们坚信，尽管这次我们的尝试并不完美，但这种勇敢的尝试为我们提供了宝贵的学习和经验，这将有助于我们在未来的工作中做得更好。

同时，我们也意识到，对于一个模型来说，训练时间的长短并不是唯一的决定性因素。还有其他重要的因素，如数据质量、模型架构的选择和优化等。在未来的研究中，我们将继续关注这些因素，以期提高我们的模型性能。

总的来说，虽然我们没有在这次尝试中找到最佳的参数配置，但我们从中学到了很多。我们了解了哪些参数组合在有限的时间内是可行的，哪些是不可行的。这为我们未来在更短的时间内训练出更优秀的模型提供了宝贵的参考和经验。虽然这是一次勇敢的尝试，但我们相信，只有通过尝试和失败，我们才能找到最好的解决方案。

签名：