基于机器学习的股价预测模型

第二期报告

武明睿1, 潘高翔1, 董浩林2, 时天行2

1 北京大学化学与分子工程学院

2 清华大学

# 摘要

我们选取AAPL和MSFT数据，完成了LSTM和GRU模型的搭建与训练，并使用两步法优化和TPE算法探索了模型的最优参数，表现最佳的模型在两个数据集上分别达到了MSE为0.121和0.475的结果。

# 引言

RNN(Recurrent Neural Net)是一类用于处理变长序列的神经网络模型，因其可以识别和处理序列中的自相关性质，RNN被广泛用于NLP、时间序列分析等领域。RNN的里程碑，LSTM (Long Short-Term Memory)由Hochreiter和Schmidhuber于1997年发明。LSTM具有图(1)所示的结构：

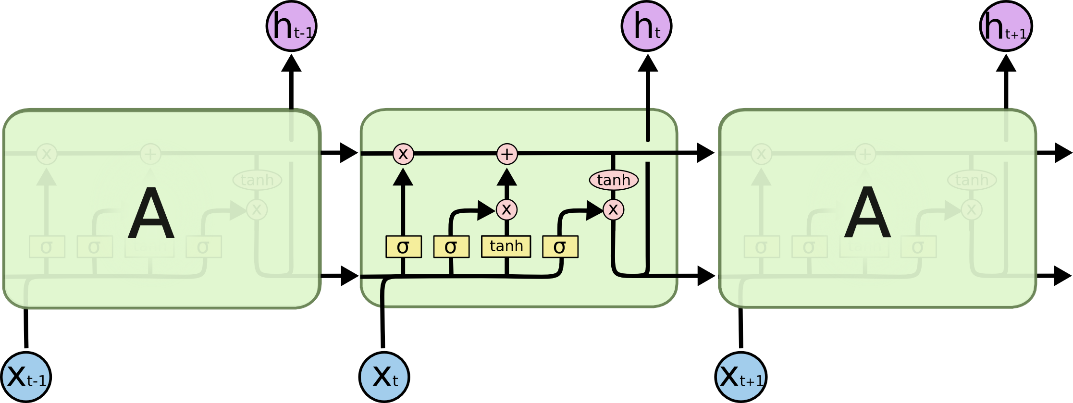


图1 LSTM网络模型

LSTM的核心是“细胞状态”向量，即方框最上方的水平直线。在一次更新中，细胞状态先后经过“遗忘门”、“记忆门”、“输出门”，从而实现信息的选择性丢弃、更新和输出。相对于传统RNN设计，LSTM的网络结构十分有利于远期信息留存(“长期记忆”)。2014年，Cho, et al. (2014)在LSTM的基础上发明了GRU(Gated Recurrent Unit)(图(2))，GRU将LSTM的遗忘门和记忆门合并为一个“更新门”，同时合并了细胞状态和隐状态向量。在实际应用中，GRU具有更少的参数和更快的训练和收敛速度，同时维持着与LSTM相当的性能。

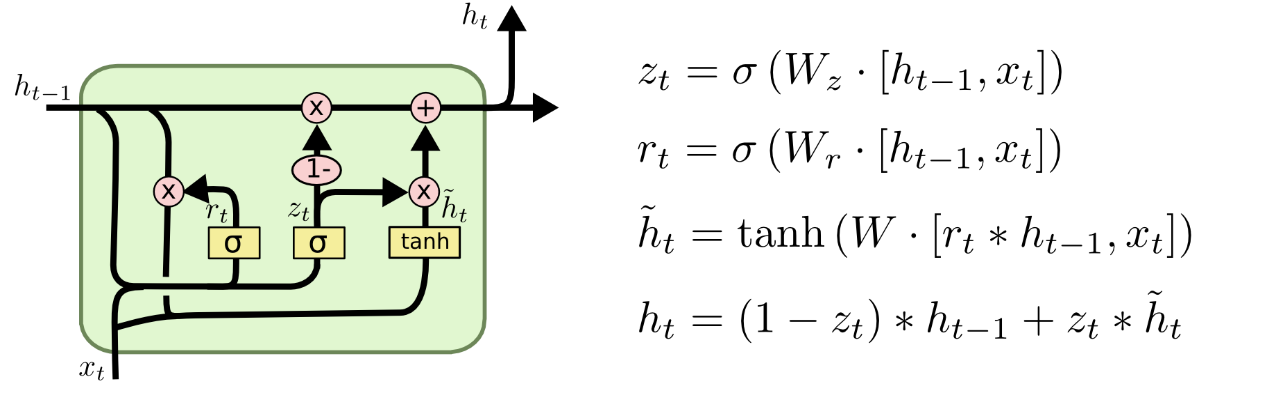


图2 GRU网络模型

我们小组分别搭建了LSTM和GRU模型，并在苹果和微软公司的月度股价数据集上训练和测试。

# 方法

## 数据集选择与预处理

我们首先从[www.quandl.com](http://www.quandl.com)下载了IBM、INTC、MSFT、TSLA和AAPL五个公司自1990年至2017年的月度股票数据，并进行差分。VR Test结果如下：

表1 五家公司股价的VR test结果

|  |  |
| --- | --- |
| Company | P value |
| IBM | 0.21 |
| INTC | 0.43 |
| MSFT | 0.19 |
| TSLA | 0.96 |
| AAPL | 0.08 |

可见五支股票的价格都没能在5%的显著水平上推翻原假设。为了尽可能保证可预测性，我们选取了p value最小的两支股票——AAPL和MSFT进行分析。

我们首先将两支股票数据的后30%作为测试集，用于评估模型表现。训练集数据归一化为标准正态分布，随后依序截取长度为 的短序列训练集，短序列最后一个数据作为标签，之前的数据作为训练输入。将训练集打乱后，从中随机抽取20%作为验证集用于超参数调节参照。

## 模型搭建与训练

网络模型由RNN单元和一个线性变换组成。 期的数据依次输入RNN单元，随后经过一个线性全连接层，得到一维标量输出作为下一期的预测。LSTM模型与GRU模型的输入输出形式基本一致，必要的参数分别为隐状态维度 和网络层数 。

使用自助法在训练集上采样，单步训练样本大小为 ，共训练 轮，观察是否收敛。训练中使用Adam优化算法，学习率为 ，优化目标为最小化MSE(Mean Square Error)。在AAPL数据集上初步训练曲线如图3所示。

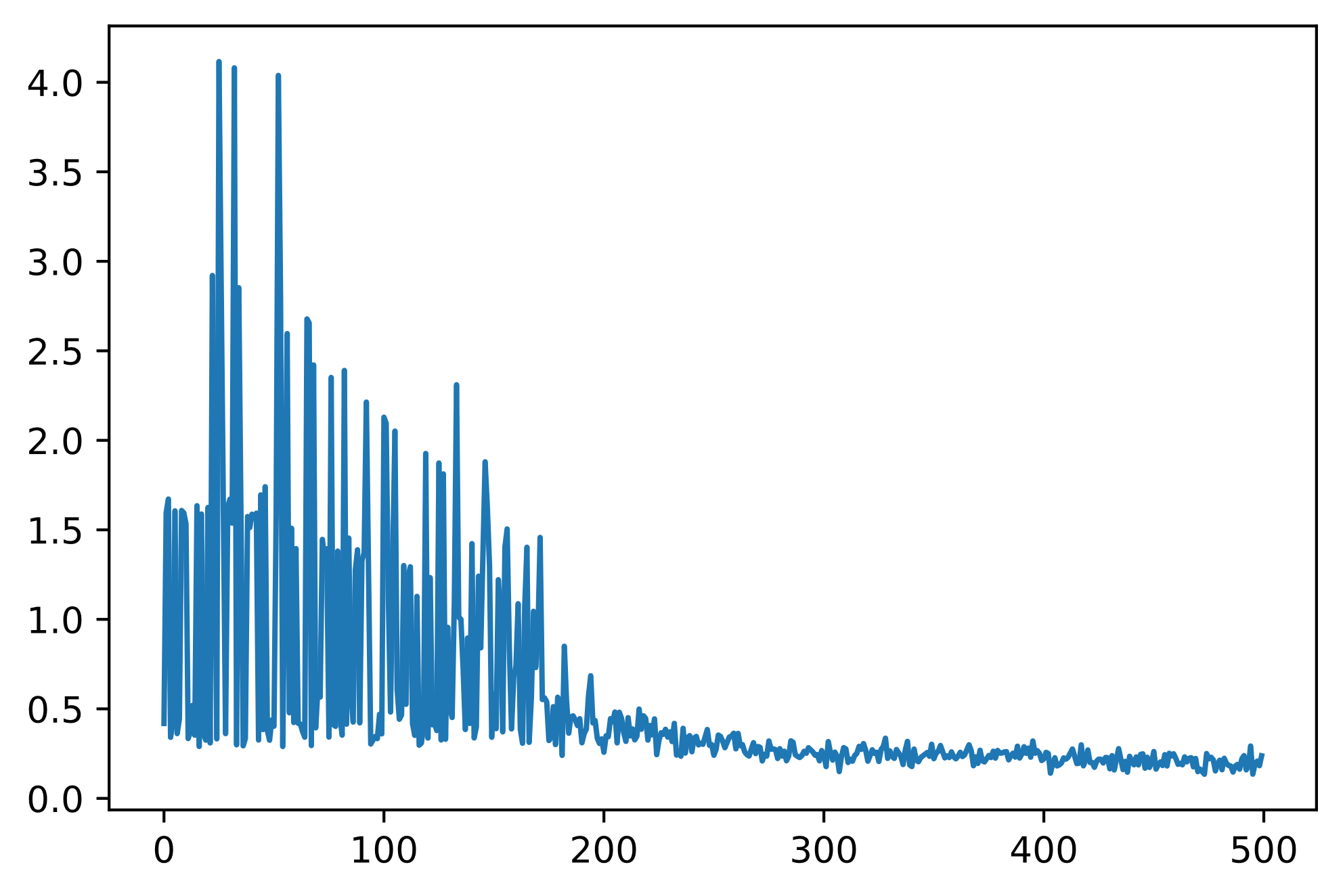


图3 训练轮数—MSE得分曲线。训练用时2.3 s (NVIDA TESLA V100GPU)

MSE在训练200轮后收敛于0.3附近。

## 超参数优化

如图3标注，模型需要优化的参数共有6个。超参数优化的目标为最小化验证集上的MSE。由于训练中模型表现波动大，我们在训练的最后10轮进行了10次评估，取均值以减小表现波动的影响。

若采用随机搜索算法，探索空间过于庞大且效率低下。为缩小搜索空间，我们假设在训练收敛的情况下，模型参数与训练参数对模型表现的影响相互独立，因此我们可以通过两步法优化得到最优参数。第一步，根据初步训练判断，的情况下基本可以达到收敛，因此固定训练参数，对模型参数进搜索；第二步，取表现最佳的模型参数，对训练参数进行搜索。采用Tree of Parzen Estimators (TPE) 黑盒优化算法进一步提高探索效率。优化参数与验证集表现见图4。

所有最优MSE均小于1，低于数据方差(数据经过归一化，方差为1)。GRU与LSTM的表现非常接近。

## 模型评估

使用最优参数和全部训练集数据(不划验证集)训练模型，将模型滚动预测(上一期的预测作为下一期的输入)与测试集进行比较，得到测试集上的MSE。

我们评估了参数优化过程中表现最好的两个模型：AAPL-GRU(图5)和LSTM-GRU(图6)。



图4 超参数优化结果

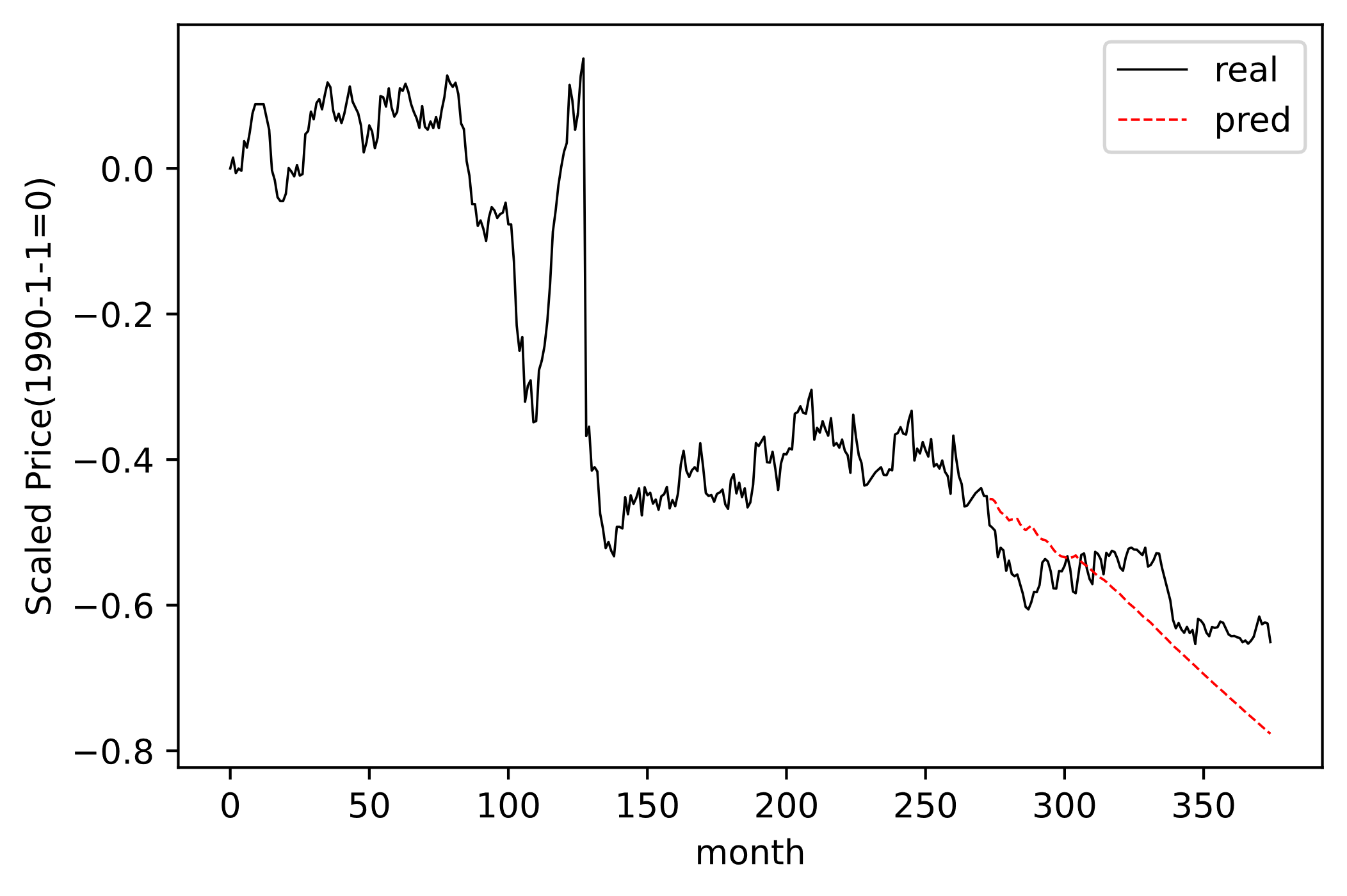


图5 AAPL-GRU测试集表现(积分曲线)

MSE = 0.121 (差分数据)

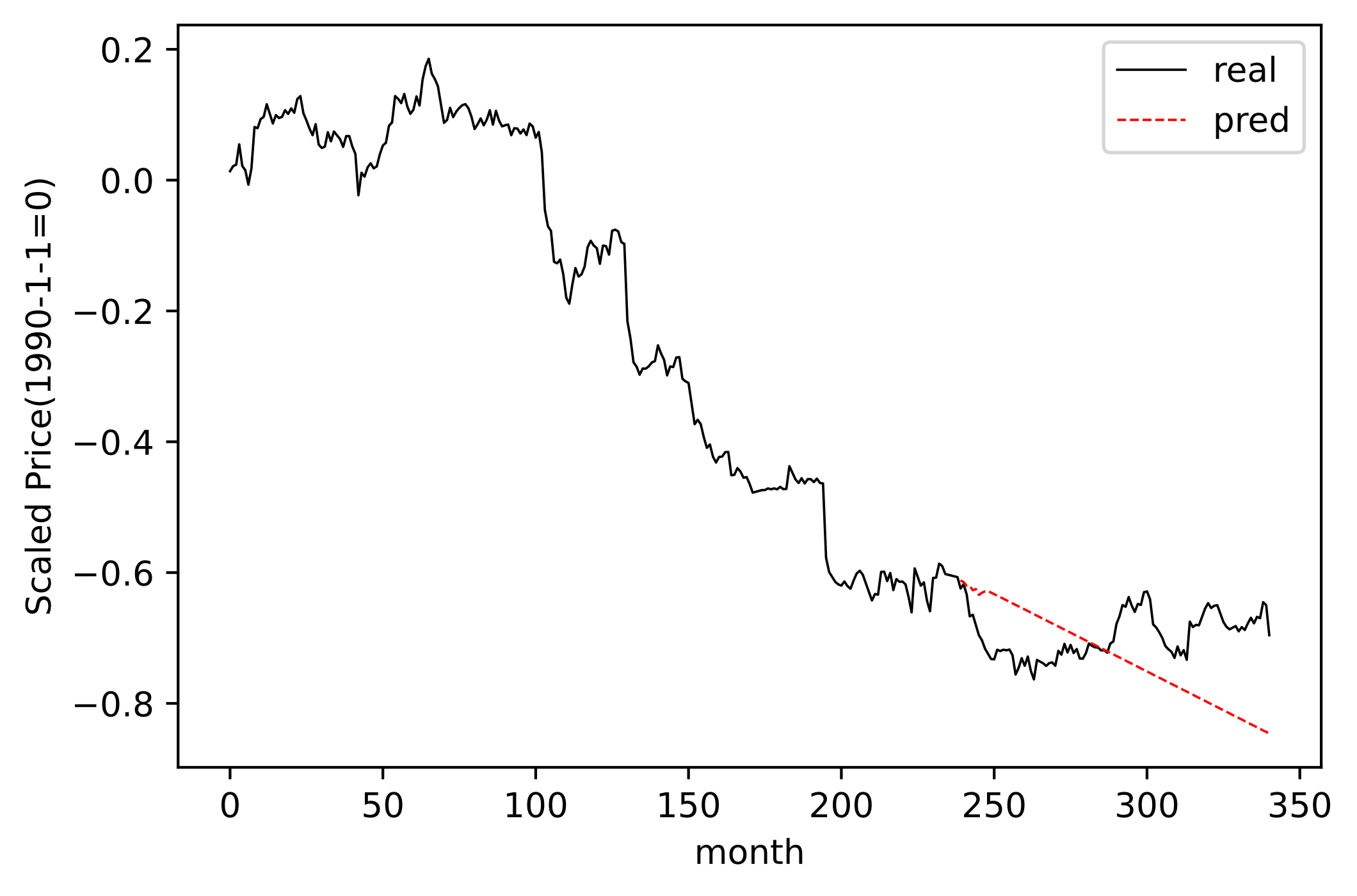


图6 MSFT-GRU 测试集表现(积分曲线)

MSE = 0.475 (差分数据)

由于滚动预测的性质，误差会随着预测期限延长而累加，造成后面的预测可能完全不正确。与移动平均或自回归方法类似，RNN方法的预测也出现了平滑现象。另一个有趣的现象是，按照最优的训练轮数进行训练，模型都无法收敛。这可能是因为模型在没有收敛的时候波动大，容易出现在验证集上MSE很小的情况；而收敛后出现过拟合，在验证集上MSE会稍微变大，所以调参过程更容易选中不收敛的参数。这个问题是十分致命的。

# 未来安排

为了提供对照，我们计划使用ARIMA模型进行相同的预测，并对比二者的表现差异。此外，以上所训练的RNN模型仍有瑕疵，主要表现为训练不收敛。将优化目标调整为训练末端的上置信边界，或不优化训练轮数和batch\_size可以解决这一问题。此外，MSE是绝对误差量，不能很好地反映测试精度，相对误差或R2 score更合适。