用于股票走势预测的时空 Transformer

Daniel Boyle 和 Jugal Kalita 科罗拉多大学科罗拉多斯普林斯分校 1420 Austin Bluffs Pkwy 科罗拉多斯普林斯，CO 80919

arXiv:2305.03835v1 [cs.LG] 5 May 2023

摘要

金融市场十分引人入胜，只要时机得当，投资者便有机会获得丰厚利润。不幸的是，金融市场具有动态、非线性的特性，因此极难预测未来的价格走势。在美国股票交易所，影响公司股票价格的因素数不胜数，包括但不限于财务报表、社交媒体和新闻情绪、整体市场情绪、政治事件以及交易心理。人类几乎不可能将这些因素关联起来。因此，我们提出了 STST，这是一种采用时空 Transformer-长短期记忆网络模型进行股票走势预测的新方法。我们的模型在 ACL18 和 KDD17 数据集上分别取得了 63.707% 和 56.879% 的准确率。此外，我们还通过模拟检验了该模型在现实中的适用性。其收益至少比标准普尔500股票指数高出 10.41%，年化收益率至少为 31.24%。

引言

股票市场是一个公开的金融市场，使公司能够出售自身股份，以便从投资者那里募集所需资本。从投资者的角度来看，股票市场提供了一个买卖单家公司或一组公司股份的场所；持有的股份数量决定了投资者拥有的部分所有权。排除外部因素的影响，这些股份的价格会因供求关系而波动。投资者通常会在认为股价将会上涨时买入股票，从投资中获利。随后，当投资者认为股价将会下跌时，便会卖出股票，以避免未来蒙受损失。

这造就了一种十分引人入胜的环境，投资者有可能从中获得丰厚利润。美国证券交易所每天的交易额都高达数十亿美元（Dunne 2015）。每笔交易背后，投资者都希望准确预测价格波动，以获取利润。遗憾的是，由于股票市场具有动态、非线性的特性，预测股价何时上涨或下跌是一项极其复杂的任务。除了公司基本面和股票价格走势之外，还有许多因素会增加预测难度，包括公众舆论、政治形势、交易心理、媒体情绪，甚至当前是星期几（Maberly 1995）。

由于交易的复杂性及其潜在盈利能力，人们投入了大量研究来预测价格走势和趋势。传统上，金融市场预测主要采用两种分析方法：技术分析和基本面分析。技术分析的理念是，价格与成交量数据包含了准确预测市场所需的全部信息。因此，它着重研究价格走势和图表形态，以帮助识别趋势、反转形态及其他类型的技术信号（Edwards、Magee 和 Bassetti 2018）。另一方面，基本面分析侧重于通过分析公司的财务报表以及各种微观与宏观经济因素，衡量证券的内在价值（Abarbanell 和 Bushee 1997）。投资者通常使用这种分析方法寻找定价错误的证券，以确定投资方向。尽管许多投资者和专家都坚定支持其中一种分析方法，但也存在许多结合两种方法的子方法，以便更全面地了解股票市场的行为。

近年来，机器学习在数据处理与分析领域取得了显著进展，在复杂的预测与回归任务中应用时也获得了令人鼓舞的成果。这引起了金融专家和分析师的关注，他们开始将其应用于金融市场。近期研究已将支持向量机（SVM）（Hearst 等，1998）、提升树（Chen 和 Guestrin，2016）以及复杂的神经网络模型等机器学习方法用于金融市场领域的不同任务。这些方法之所以广受欢迎，是因为它们天生具备处理海量数据并提取复杂特征的能力，而这些特征原本难以获取。机器学习的主要应用方向包括交易信号预测（Wu 等，2020）、价格预测（Niaki 和 Hoseinzade，2013；Sen 和 Chaudhuri，2017）以及价格走势预测（Li、Dai 和 Zheng，2022）。

版权所有 © 2022，人工智能促进协会（www.aaai.org）。保留所有权利。

交易信号预测旨在判断买入和卖出某项证券的时间或价格，以实现利润最大化、损失最小化。这些时点通常称为“信号”。价格预测则旨在预测未来有限个时间步内某项证券的收盘价或移动平均值。最后，价格走势预测侧重于预测指定时间范围内价格走势的方向。

对于日内交易者而言，最常见的目标是准确预测股票价格何时可能上涨或下跌，从而从价格走势中获利。由于影响股票市场的外部因素众多，而且在任何给定时刻，针对特定证券都可能有大量数据可用，因此，日内交易者极难从数据中识别出所有必要的模式，以准确预测未来的价格走势。近年来，Transformer模型凭借自注意力机制学习数据中复杂的长程时间关系的能力，在各种序列相关问题中广受欢迎。基于这一理念，我们提出了面向股票的时空Transformer（STST），这是一种专注于关联时空模式、用于股票走势预测任务的Transformer-LSTM模型。

背景

传统上，长短期记忆网络（LSTM）（Hochreiter 和 Schmidhuber，1997）和卷积神经网络（CNN）（Hochreiter 和 Schmidhuber，1997）一直是序列任务中最常用的模型。尽管这两种模型至今仍非常有效，但也各有不足。对于长短期记忆网络，数据会跨越时间步边界按顺序输入，这导致训练过程极其缓慢。处理大量数据时，这会成为一个严重问题。另一方面，卷积神经网络一次性接收整个序列，因此训练速度快得多。但遗憾的是，由于卷积神经网络一次性处理整个序列，它们无法保留位置信息。此外，它们在高效学习长期依赖关系方面也可能表现不佳（Linsley 等，2018）。增大感受野可以克服这一问题，但计算开销很快会变得非常大。

Transformer

为克服其中许多挑战，Vaswani 等人于2017年在 Google 提出了用于语言序列任务的 Transformer 神经网络架构（Vaswani 等人（2017））。他们的方法表明，采用自注意力的序列模型在序列到序列（seq2seq）任务（如语言翻译）上能够达到最先进水平，而此类任务此前一直由长短期记忆网络主导。原始 Transformer 架构见下方图1。

与以往用于序列到序列任务的基于 RNN 的架构类似，Transformer 由编码器和解码器组成，每个部分都包含可堆叠的注意力机制层和前馈神经网

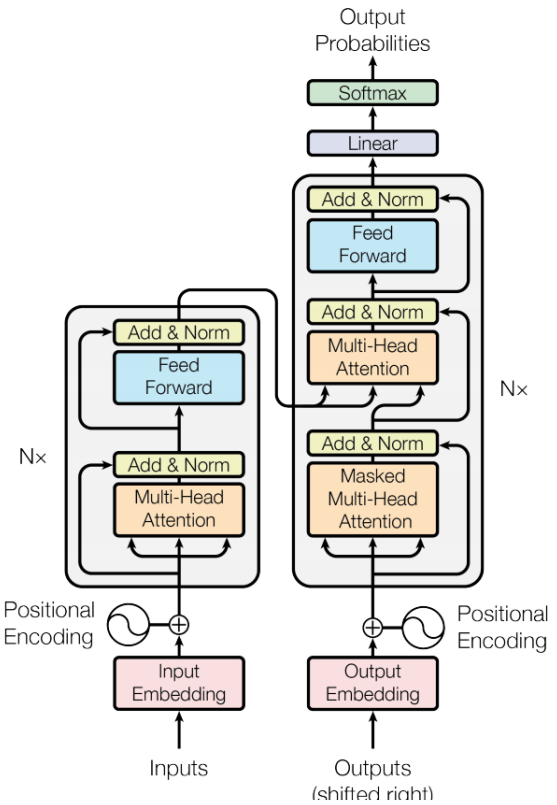


图1：原始Transformer架构（Vaswani 等人（2017））

络。在 Transformer 中，有五个关键组件：位置编码、多头注意力、前馈网络、残差连接和归一化。

位置编码 在传统的序列到序列模型中，循环神经网络会逐个输入元素，从而为序列中的词元形成位置概念。对于 Transformer，整个输入会一次性馈入网络；因此，需要在输入特征向量中加入某些信息，以编码位置信息。这项任务由一个称为位置编码器的层来完成。在原始 Transformer 中，使用固定的正弦函数生成位置编码，并将其与原始嵌入输入相加。近年来，位置嵌入层日益流行，它们使用可训练模型进行位置编码。这种做法见于 RoBERTa（Liu等人，2019年）和 GPT-2（Radford等人，2019年）等知名的 Transformer 语言模型。

多头自注意力 可以说，Transformer架构中最重要的组件就是多头自注意力。自注意力是一种机制，用于建立序列中不同变量之间的重要性关系，从而帮助模型判断应该关注什么。多头自注意力允许多个不同的自注意力组件（注意力头），以便

学习序列中的不同关系。这样，Transformer便能够学习复杂的长期依赖关系，同时避免循环神经网络面临的许多问题。

前馈网络与其他神经网络架构类似，Transformer 会在编码器和解码器中使用前馈稠密神经网络，这些网络位于多头注意力层之后。这些层将多头注意力层的输出映射到所需的输出向量空间。通常还会在网络输出之后使用修正线性单元（ReLU）激活函数（Nair 和 Hinton，2010）。

残差连接由于 Transformer 往往规模极大，因此容易出现梯度消失问题。解决这一问题的常见方法是在网络中添加残差连接。残差连接本质上是一种让数据跳过某些层，并将其添加或拼接到后续某一层输出上的方式（He 等人，2016）。这有助于梯度在反向传播过程中进一步向前传递，从而支持构建规模大得多的网络。

归一化Transformer架构中最后一个重要层是归一化层。归一化层传统上用于帮助稳定神经网络的动态特性。原始 Transformer 模型采用了一种称为层归一化的归一化方法，重点是对神经网络中每个维度上的层输入进行归一化。相比批归一化（大型卷积神经网络中常见的一种技术），它被优先选用，因为自然语言处理（NLP）任务中的词嵌入通常在各个维度之间存在较高方差（Vaswani 等人（2017））。

相关工作

在接下来的章节中，我们将介绍利用 Transformer 建模复杂关系的相关工作，以及将 Transformer 用于时间序列数据任务的研究。我们还将回顾股票走势预测领域的已有研究。

基于 Transformer 的编码 基于编码器的 Transformer 近年来日益流行，成为一种高效学习多种不同类型数据潜在表示的方法。在表示学习领域展现出巨大潜力的最著名的基于 Transformer 的模型之一，是一种名为 BERT（即 Transformer 双向编码器表示）的模型，由 Google 于 2019 年推出（Devlin 等人，2019）。BERT 是一种基于Transformer的编码器方法，用于学习词元和句子的上下文化双向嵌入。该方法引入了两种新颖的无监督预训练技术，以充分实现词元或句子的上下文化。第一种方法称为掩码语言建模（MLM）。MLM 通过遮盖输入中的随机词元或词元组，对编码器进行预训练，并利用模型预测缺失的词元。其次，他们引入了下一句预测（NSP）预训练技术，重点在于预测某个特定句子是否紧跟在另一个句子之后。BERT 在 11 项不同的自然语言处理相关任务中取得了最先进水平的结果。

转向时间序列领域，2021 年，Zerveas 等人提出了一种基于 Transformer 编码器的模型，用于对多变量时间序列进行无监督表示学习（Zerveas 等人，2021）。他们的方法采用了与 BERT 的 MLM 预训练方法类似的策略：随机屏蔽序列中的变量，并让模型尝试预测这些变量。该方法在多个不同的多变量时间序列数据集上的分类和回归任务中，均超越了当时最先进的模型。

此外，Dang 等人提出了一种名为 TS-BERT 的时间序列 BERT 方法，用于检测单变量时间序列数据中的异常（Dang 等人，2021）。与 Zerveas 等人使用嵌入层对时间序列进行编码不同，TS-BERT 使用时间序列字典创建聚合序列，再将其转换后用于传统的 BERT 模型。TS-BERT 在时间序列异常检测方面超越了当时最先进的水平。

股票走势预测预测股票未来走势方向的能力通常被称为“股票走势预测”。由于这一问题本身复杂，且潜在收益丰厚，它一直吸引着投资者和研究人员。大多数股票市场预测方法可分为两类：一类仅使用价格数据，另一类则将价格数据与其他外部信息相结合。此类信息包括新闻文章或社交媒体帖子等。本文仅关注只使用历史价格数据的相关研究。

2018 年，Feng 等人提出了 Adv-ALSTM，这是一种基于时间注意力机制的长短期记忆网络（LSTM）（Feng 等，2018），其基础是 Qin 等人的 ALSTM 模型（Qin 等，2017）。他们通过添加扰动进行对抗训练，以模拟股票市场价格数据的随机性。与当时最先进的模型 StockNet（Xu 和 Cohen，2018）相比，他们的方法取得了更高的准确率和马修斯相关系数得分。该模型在 ACL18 和 KDD17 数据集上进行了评估，这两个数据集常用于价格走势预测任务，其准确率分别为 57.20% 和 53.05%。

Zhang 等人在 2021 年提出了一种名为 FA-CNN 的方法，该方法结合了名为 DeepFM（Guo 等，2017）的深度因子分解模型与基于注意力机制的 CNN（ATT-CNN）。他们利用 DeepFM 模型提取日内空间特征，同时采用 ATT-CNN 模型提取不同时间点之间的时间特征（Zhang、Liu 和 Zheng，2021）。他们在中国股票市场的食品饮料、医疗保健和房地产板块的三个数据集上应用了该模型。该模型在每个数据集上都取得了最先进水平的结果。

不久之后，Yoo 等人提出了一种名为数据轴 Transformer 多级上下文（DTML）（Yoo 等，2021）的模型，用于预测股票走

势。该方法将多只个股和全球市场指数输入带有时间注意力机制的长短期记忆网络，为每个输入捕获一个上下文向量。随后，将各只股票和全球市场指数对应的上下文向量堆叠起来，为当前市场状态提供多层级上下文。接着，将结果输入 Transformer编码器，以关联不同的上下文并生成整体预测。他们使用三个国外数据集和三个美国数据集对模型进行了测试，其中包括 ACL18 和 KDD17数据集。模型的准确率分别达到 57.44% 和 53.52%，超越了当时的最先进水平。

2022年初，Li等人提出了一种选择性迁移学习方法 STLAT，该方法将带有时间注意力的长短期记忆网络与对抗训练相结合，用于股票走势预测（Li、Dai 和 Zheng，2022）。通过有选择地考量源数据中哪些部分有用，并将其与对抗训练相结合，该方法使他们能够在基础注意力 LSTM模型的基础上进一步提高准确率。他们在 ACL18和 KDD17数据集上测试了该模型，分别取得了 64.56% 和 58.27% 的准确率，达到当前最先进水平。

随后，在2022年中，He 等人提出了基于实例的注意力深度迁移学习（IDTLA）（He、Siu 和 Si，2022）。他们的方法会根据更大的数据集，选择与训练最相关的源数据。这样一来，即使可用价格数据较少，也能有效地针对这些股票训练模型。为了与最先进的方法进行比较，他们在 ACL18数据集的基于行业的子集上训练模型。通过对各个数据集上的结果取平均值，他们得到的总体准确率为 58.2%，马修斯相关系数得分为 0.168。虽然该方法未能超越最先进水平，但他们提出了一种新颖的方法，能够针对历史数据极少的股票训练出准确率相对较高的模型。

问题定义

股票走势预测的目标是学习一个函数 $f(\hat{x},\ \theta,\ T) \rightarrow \{0,1\}$，其中 $\hat{x}$ 表示输入向量 $<x_1,\ x_2,\ ...\ x_t>$，$x_i \in \mathbb{R}$表示模型参数，θ 表示可配置的时间范围，T 表示价格变动方向。因此，给定某只股票的输入 $\hat{x}$，目标是预测经过 T 个时间步后，股价会上涨（1）还是下跌（0）。我们的研究目标以及前文提及的相关工作，都着眼于时间范围 $T = 1$。也就是说，我们关注的是预测股价在下一个交易日会上涨还是下跌。

方法

算法交易和技术交易的大部分工作，都涉及从历史价格数据和技术指标中识别关键特征与模式。这些模式可能彼此相近，也可能相隔数周。因此，我们的模型必须能够成功发现并捕捉这些模式，为此，我们选择使用 Transformer，因为它能够借助自注意力捕捉复杂的长期依赖关系。接下来的章节中，我们将介绍时空股票 Transformer（STST），这是一种用于进行股票走势预测的 Transformer-LSTM 模型。

模型

我们的模型采用基于Transformer的编码器，在指定的上下文窗口 W 内捕获个股的上下文化嵌入。这些上下文化嵌入被输入一个多层长短期记忆网络，并与稠密神经网络结合，以生成最终预测。为了提升模型性能，我们对传统 Transformer的若干部分进行了修改，包括位置编码和嵌入层。STST模型如图3所示。

位置编码 我们没有采用传统的正弦位置编码或较新的可训练位置编码器，而是选择使用 Date2Vec。Date2Vec 是一种基于时间的编码器 Time2Vec 的预训练变体，已被证明能够提升基于时间序列的预测任务的结果（Kazemi等人，2019年）。与正弦位置编码层和学习式位置编码层类似，Time2Vec 旨在创建正交的时间嵌入。这样一来，就可以将位置向量添加到特征向量中，而不改变输入特征向量的原始含义。

时空嵌入与自注意力 目前最先进的股票预测方法都采用了时间注意力机制。这很有用，因为它能帮助模型学会关注特定的时间事件，从而做出更好的预测。然而，时间注意力机制的缺点是，它只关注如何建立序列中不同时间步之间的关系。虽然这很有用，却没有考虑不同特征之间可能存在的空间相关性，也没有考虑同一时间步内可能存在的其他关系。例如，单个时间步内开盘价、收盘价、最高价和最低价之间的差异，能够提供有助于日内交易者做出决策的信息。这些情况存在于空间边界上，传统的时间注意力机制难以捕捉。

为此，我们创建了一个特殊的时空嵌入层，将输入特征转换为展平的特征向量。下面将详细介绍这一层。设每个输入表示为 $X \in \mathbb{R}^{N\times(T+F)}$，其中$N$为上下文窗口大小，$F$为非时间特征的数量，$T$为时间特征的数量。时空嵌入将 X 拆分为 $X_f \in \mathbb{R}^{N\times F}$和 $X_t \in \mathbb{R}^{N\times T}$。随后，对 $P_t = Time2Vec(X_t)$应用位置编码。接着，将特征矩阵 X_f展平为 X_f'，使得 $X_f' \in \mathbb{R}^{1\times(NF)}$。展平后，将其与 X_t 拼接，得到结果 X_c，其中 $X_c = \{x_{fi}' \oplus x_{ti} | x_{fi}' \in X_f', x_{ti} \in X_t\}$。拼接完成后，将结果向量输入输出大小为 D 的密集神经网络，以得到 $E_X = Dense(X_c)$，

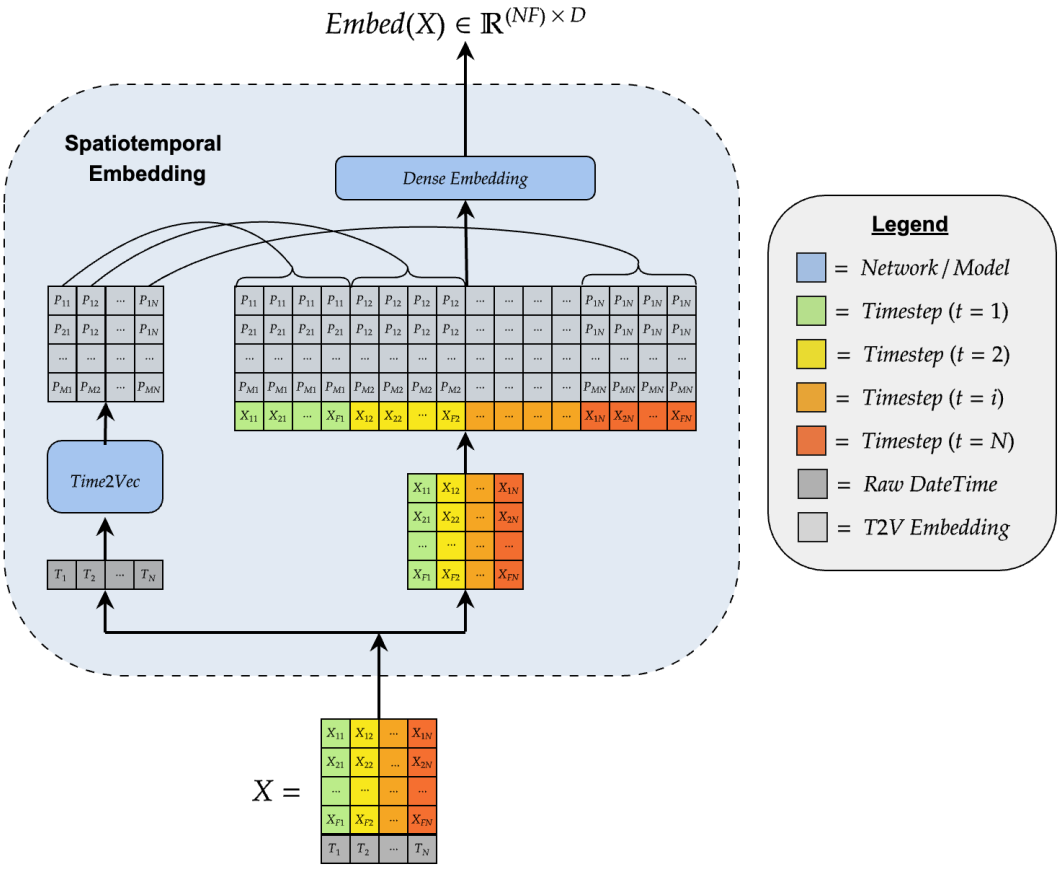


图 2：时空嵌入

其中 $E_X \in \mathbb{R}^{NF \times D}$。图 2 展示了这一过程的视觉表示。

输入 X 转换为 E_X后，会被传递到 Transformer 编码器，其中包含一个传统的多头自注意力模块，这种模块常见于原始 Transformer 中（Vaswani 等人（2017））。生成的自注意力图维度为 $NF \times NF$。这使多头注意力模块能够捕捉空间和时间边界上各个特征之间的关系。为进行评估，我们将使用这种时空嵌入与使用传统嵌入的模型版本进行比较，以观察其对模型分类结果的影响。

编码器 与原始 Transformer 类似，我们的编码器由数量可配置的多个模块组成，每个模块包含多头注意力机制、归一化、前馈神经网络和残差连接。不过，原始 Transformer 使用层归一化，而这种方法通常无法在涉及时间序列数据的任务中取得良好效果。因此，我们选择使用幂归一化，这是一种重新设计的批归一化形式，能够克服批归一化带来的许多问题（Shen 等人（2020））。在模型参数搜索过程中，我们测试了不同类型的归一化，以确定哪种方法能为我们的问题带来最佳结

果。我们发现，相较于批归一化或层归一化，幂归一化始终能带来更稳定的结果。除了归一化类型外，我们还研究了归一化的最佳位置。原始 Transformer 会在自注意力和位置前馈网络之后立即进行归一化。不过，有研究表明，在这些层之前进行归一化可能会带来更好的结果（Xiong 等人（2020））。我们的模型搜索发现，传统的后置归一化最适合我们的需求。

经过一系列 Transformer 编码器块后，由于嵌入层中空间特征的展开，编码器输出 Y_E 的形状为 $NF \times D$。然而，我们希望分类层之前的最终输出在上下文窗口中的每个时间步都包含一个长度为 C 的单一上下文向量。因此，我们将 Y_E 中的空间特征重新堆叠为一个形状为$N \times C$ 的新输出 Y_{E2}，其中 $C \in \mathbb{R}^{FD}$。

分类器 我们比较并测试了模型的两个不同变体，每个变体都在 Transformer 编码器之上采用了略有不同的分类模型。我们测试的第一个模型采用两层前馈神经网络。第二个模型采用多层长短期记忆网络，后接单层前馈神经网络。长短期记忆网络

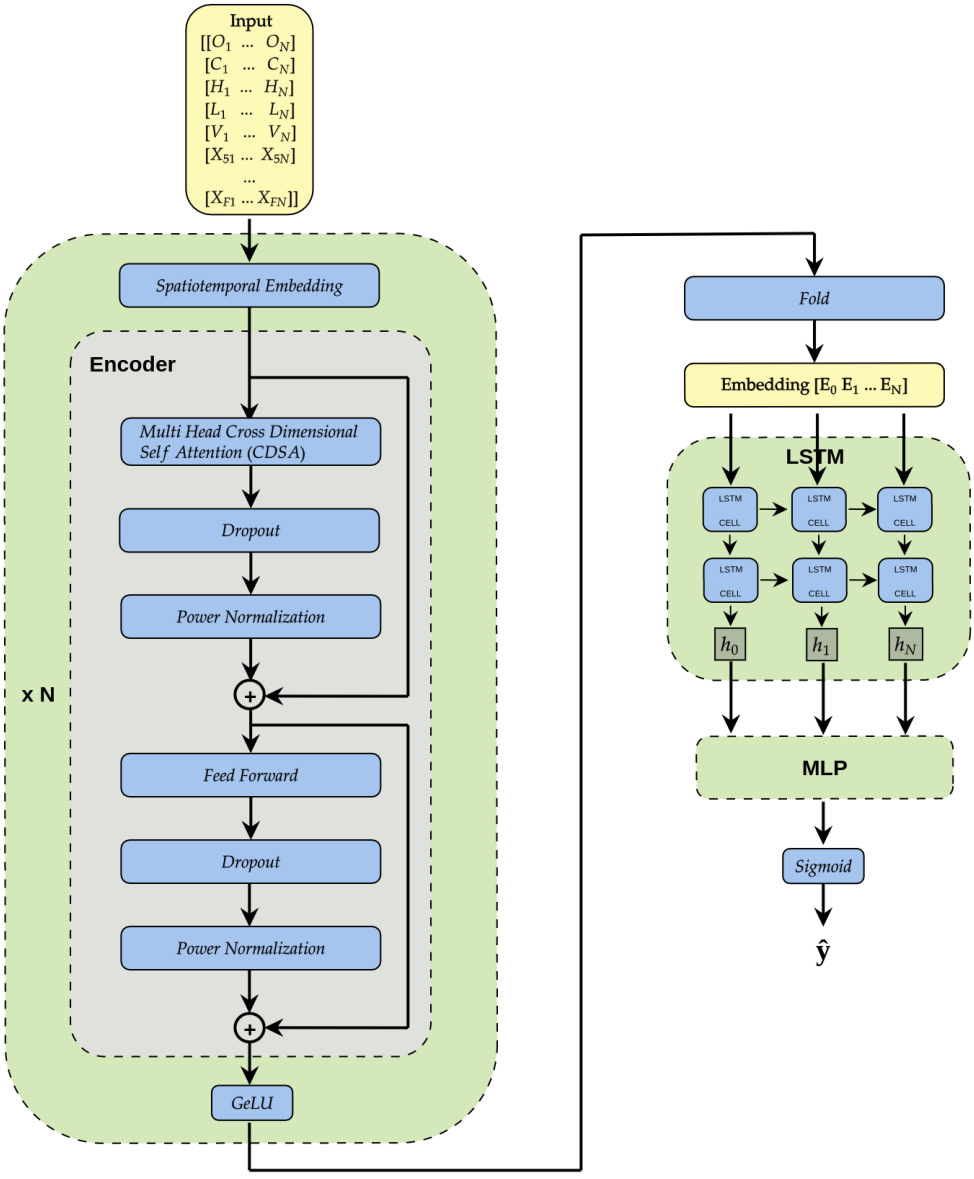


图 3：STST：时空股票 Transformer

已在使用原始特征进行股票市场预测方面展现出良好的效果；此前最先进的方法均使用了这些特征（Feng 等，2018）（Yoo 等，2021）（Li、Dai 和 Zheng，2022）。不过，我们并未将原始特征直接输入长短期记忆网络，而是输入由 Transformer 编码器输出的编码后上下文特征。由于价格走势预测是一个二分类问题，每个分类模型的输出都是一个单一数值。该数值经过 Sigmoid 激活函数处理后，得到输出预测值 $\hat{y} \in [0, 1]$.

训练 STST模型训练 100 轮，并采用提前停止来防止模型过拟合。为进一步改进训练，我们采用学习率预热阶段，在指定的步数内将学习率从 0 线性提高至最大值，λ。预热是提高 Transformer 在训练过程中稳定性的常见方法（Popel 和 Bojar，2018）。

为了调整模型，我们采用基于网格的超参数搜索，以确定最佳参数组合。每个超参数的搜索空间

参数	搜索空间
学习率	[1e-7 到 1e-4]
批大小	[16, 32, 64, 128]
预热步数	[0:50000,5000]
上下文窗口大小	[4,8,16,32,64]
编码器数量	[1,2,4,8]
注意力头数	[1,2,4,8]
模型维度	[32,64,128,256]
d_ff	[256,512,1024, 2048]
前馈网络丢弃率	[0 至 0.5]
注意力丢弃率	[0 至 0.5]
LSTM层数	[1,2,4]
LSTM隐藏层大小	[128,256,512]
归一化类型	[批次、层、幂]
归一化位置	[前置、后置]

表1：STST模型的超参数选择。

见表1。由于模型较为复杂，需要调整多种超参数，以确保获得最佳性能。超参数可分为三类：训练参数、尺寸参数和架构参数。我们重点调整的训练超参数包括学习率、批大小、所需的预热步数，以及注意力机制和前馈网络的丢弃率（注意力丢弃率、前馈网络丢弃率）。接下来是尺寸参数，这些参数用于设定模型各个部分的大小。我们对多种尺寸参数进行微调，包括输入上下文窗口大小、编码器模块数量（n encoders）、注意力头数（n heads）、Transformer的编码输出特征向量大小（模型维度，d model）、每个全连接层的节点数（前馈层维度，d ff）、LSTM层数（n lstm layers）以及LSTM网络的隐藏层大小（LSTM隐藏层大小，d lstm hidden）。最后一类需要调整的参数是架构参数。第一个架构参数是所用归一化层的类型。对此，我们测试了批归一化、层归一化和幂归一化。最后一个架构参数是归一化位置。我们会确定模型在Transformer编码器内部的自注意力和位置前馈网络之前（pre）还是之后（post）进行归一化时性能最佳。

Data

数据集

为了测试我们模型的性能，我们选择了 ACL18（Xu 和 Cohen，2018）和 KDD17（Zhang、Aggarwal 和 Qi，2017）数据集。选择这些数据集是因为它们公开可用，且常用于比较模型在股票价格走势预测方面的表现。

ACL18数据集包含来自美国纳斯达克和纽约证券交易所88只高成交量股票的数据，截取自-

变化。它包含从2014年1月1日至2016年1月1日的历史价格数据。在数据划分方面，ACL18数据集进一步划分为：训练集的时间范围为2014年1月1日至2015年8月8日；验证集的时间范围为2015年8月8日至2015年10月1日；测试集的时间范围为2015年10月1日至2016年1月1日。

对于 KDD17，该数据集包含来自纳斯达克和纽约证券交易所成交量最高的 50 只股票的数据。KDD数据集包含从 2007年1月1日 到 2017年1月1日 的历史价格数据。数据划分方面，KDD17数据集进一步拆分为：训练集的时间范围为 2007年1月1日 至 2015年1月1日；验证集的时间范围为 2015年1月1日 至 2016年1月1日；测试集的时间范围为 2016年1月1日 至 2017年1月1日。

两个数据集均由历史价格逐笔数据构成，包括所提供日期范围内每个交易日（不包括周末）的开盘价、收盘价、调整后收盘价、最高价、最低价以及交易量。除了提供的基础历史价格数据外，我们还基于价格数据生成额外的汇总数据，这些数据通常称为技术指标。其中包括从简单移动平均线到更复杂的动量和波动率计算等各种指标。对于我们的方法，我们额外生成了 22 个特征。前四个特征通过将日期拆分为日、月、年和工作日生成。这些特征被输入 Time2Vec 模型，用于对每个时间步内的位置进行编码。对于其余特征，我们使用基于常见技术指标计算出的 18 个二元信号。表 2 中详细列出了各项特征及其计算方法。

对于这两个数据集，输出标签 Y根据下文公式1中的以下阈值计算得出。所有落在 -0.5% 至 0.55% 范围内的数据点均予以忽略。选择这些阈值是为了帮助平衡数据集。这样一来，ACL18数据集中的两个类别占比分别为 50.64% 和 49.36%，KDD17数据集中的占比分别为 50.70% 和 49.30%。

$$\begin{cases} 0, & (AdjClose_{t+1}/AdjClose_t - 1) <= -0.5\% \\ 1, & (AdjClose_{t+1}/AdjClose_t - 1) >= 0.55\% \end{cases} \tag{1}$$

其他有助于股价预测的信息类型包括情感数据、外国市场数据和公司的财务报表。其他研究（Sawhney 等人，2020年）发现，这些补充信息有助于提升此类任务的结果，但这类信息往往难以大批量获取，因此大大增加了构建通用模型的难度。然而，本文的重点是仅使用历史数据来预测走势。因此，我们只关注基于价格的数据。

除了输入特征向量的维度和内容外，数据的另一个对我们的任务至关重要的方面是上下文窗口大小。上下文窗口表示模型一次能够看到的时间步数量。窗口太小，模型就可能难以

特征	描述	公式
Year	表示年份的十进制值	当前年份 / 3000
月	表示月份的十进制值	当前月份 / 12
Day	表示日期的十进制值	当前日期 / 31
工作日	表示星期几的十进制值	当前工作日 / 7
Open	当日开盘价	-
关闭	当日收盘价	-
调整收盘价	经过拆股调整的当日收盘价	-
High	当日最高值	-
Low	当日最低值	-
成交量	当日成交量	-
SIG简单移动平均线（10、30、50、200）	基于 N 日简单移动平均线的二元信号 （简单移动平均线）适用于 N = 10、30、50 和 200。	$\begin{cases} 1, & Close > SMA \\ 0, & otherwise \end{cases}$
SIG EMA（10、30、50、200）	基于 N 日指数移动 平均线（EMA），适用于 N = 10、30、50 和 200。	$\begin{cases} 1, & Close > EMA \\ 0, & otherwise \end{cases}$
SIG MOMENTUM	基于 10 日动量指标的二元信号	$\begin{cases} 1, & MOM > 0 \\ 0, & otherwise \end{cases}$
SIG STOCHRSI	基于随机相对强弱指数的二元信号 指数 (STOCHRSI) (长度=14，RSI 长度=14，k=3，d=3)	$\begin{cases} 1, & SRSI(i) \leq 25 \; or \\ & SRSI(i) > SRSI(i-1) \\ & and \; SRSI < 75 \\ 0, & otherwise \end{cases}$
信号 随机指标 D	基于随机指标 D 带的二元信号	$\begin{cases} 1, & STOD(i) > STOD(i-1) \\ 0, & otherwise \end{cases}$
SIG STOCH K	基于随机指标 K 波段的二元信号	$\begin{cases} 1, & STOK(i) > STOK(i-1) \\ 0, & otherwise \end{cases}$
SIG MACD	基于移动平均收敛发散 （MACD）指标的二元信号 （快速=12，慢速=26，信号=9）	$\begin{cases} 1, & MACD_{signal} < MACD \\ 0 & otherwise \end{cases}$,
SIG CCI	基于商品通道指数的二元信号 (CCI) 指标（14日）	$\begin{cases} 1, & CCI(i) \leq 100 \; or \\ & CCI(i) > CCI(i-1) \\ 0, & otherwise \end{cases}$
SIG MFI	基于14日资金流量指标的二元信号 (MFI) 指标	$\begin{cases} 1, & MFI(i) \leq 20 \; or \\ & MFI(i) > MFI(i-1) \\ & and \; MFI(i) < 80 \\ 0, & otherwise \end{cases}$
SIG AD	基于累积/派发的二元信号 （AD）指标	$\begin{cases} 1, & AD(i) > AD(i-1) \\ 0, & otherwise \end{cases}$
SIG OBV	基于能量潮指标（OBV）的二元信号 指标	$\begin{cases} 1, & OBV(i) > OBV(i-1) \\ 0, & otherwise \end{cases}$
SIG ROC	基于10日变化率（ROC）的二元信号 指标	$\begin{cases} 1, & ROC(i) > ROC(i-1) \\ 0, & otherwise \end{cases}$

表 2：STST 技术输入特征定义

获取足够的上下文信息，以便做出明智的决策。如果窗口太大，模型可能更难学习。对于我们的模型，上下文窗口大小是一个需要微调的超参数，用于平衡模型性能与收敛性。

评估

为了评估 STST 模型的有效性，我们提出了三个研究问题：

Q1. 市场预测：与其他最先进的方法相比，STST 在股票市场价格预测方面表现如何？**Q2. 消融研究**：移除该方法的若干关键方面后，STST 的表现如何？**Q3. 投资模拟**：将 STST 用于制定股票市场交易决策时，其表现如何？

我们将在以下各节中概述每项实验的详细信息。

Q1：模型性能

为进行基线比较，我们针对三种不同类型的方法测试了我们的模型。第一类是随机方法，具体做法是将我们模型的性能与随机分类器进行比较。这为展示我们的模型能够优于随机方法提供了依据。接下来，我们将模型与一个简单的基线模型进行比较。由于LSTM网络被广泛用于序列到序列任务及其他股票相关问题，我们采用多层LSTM作为基线模型。我们对LSTM进行微调，以针对每个目标数据集获得尽可能高的准确率。最后，我们将模型与其他常见的最先进水平的股票走势预测方法进行比较。以下是我们用于比较的各个模型：

- **Adv-LSTM** （Feng 等，2018）：基于时间注意力机制的LSTM与对抗训练相结合。
- **MAN-SF** （Sawhney 等，2020）：利用图神经网络（GNN）结合来自价格数据、社交媒体和股票相关性的混沌时间信号。
- **DTML** （Yoo 等，2021）：采用基于注意力机制的长短期记忆网络对个股进行上下文建模，再通过 Transformer 对股票与整体市场进行关联，以获得最终预测结果。
- **IDTLA** （He、Siu 和 Si，2022）：一种基于实例的深度迁移学习方法，采用基于注意力机制的长短期记忆网络，仅使用少量数据进行股票市场预测。
- **STLAT**（Li、Dai 和 Zheng，2022）：采用注意力机制的长短期记忆网络，结合对抗训练、数据筛选和迁移学习。目前处于最先进水平。
- **STLAT-AT**（Li、Dai 和 Zheng，2022）：不使用对抗训练的 STLAT 模型。
- **STLAT-TL** （Li、Dai 和 Zheng，2022）：不使用迁移学习的 STLAT 模型。
- **STST-MLP-T**：我们的基础 Transformer编码器模型，采用传统嵌入和MLP分类器。
- **STST-T**：我们的基础Transformer编码器模型，不使用时空嵌入（仅包含时间信息）。
- **STST-MLP**：不含长短期记忆网络-MLP分类器的 Transformer 编码器模型。
- **STST**：我们的Transformer编码器模型同时采用了时空嵌入和长短期记忆网络/MLP分类器。

继之前关于股票市场预测的研究之后，我们采用两个不同的指标来评估预测性能。第一个指标是准确率，用于总体说明模型正确预测了多少测试样本。此外，我们还使用马修斯相关系数（MCC），这是一种在比较股票走势预测方法时非常常用的指标。MCC指标计算实际值与预测值之间的皮尔逊积矩相关系数。当类别不平衡时，MCC有助于提供更准确的结果。对于这两个指标，得分越高表示模型性能越好。准确率和MCC的公式见下方公式2和公式3。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{3}$$

问题2：消融研究

基于上文概述的STST模型，我们开展消融研究，以说明长短期记忆网络和时空嵌入层对于整体模型性能的重要性和有效性。作为基线，我们训练的第一个模型采用传统的时间注意力机制和MLP分类器。该模型称为STST-MLP-T。接下来，我们训练一个采用长短期记忆网络-MLP分类器和传统时间注意力机制的变体模型。该模型旨在展示长短期记忆网络-MLP分类器对STST整体结果的有效性。下文将该模型称为STST-T。最后，我们测试的最后一个变体模型包含时空嵌入层，但仅搭配MLP分类器。该模型用于展示时空注意力在STST模型中的重要性。该模型称为STST-MLP。我们将这些结果与完整的STST方法进行比较，以说明同时采用时空嵌入和长短期记忆网络-MLP分类器对于模型整体性能至关重要。

Q3：投资模拟

在最后一项实验中，我们在模拟市场交易场景中使用STST模型，以确定其预测在交易中的效果。为此，我们使用 ACL18 和 KDD17数据集中的测试数据创建了两个模拟环境。模拟智能体的初始现金为 10000 美元。在每天结束时，智能体根据得分最高的五个正向预测，均衡配置其股票投资组合，以用于次日交易。如果正向预测少于五个，则将投资均匀分配给现有的正向预测。

模型	ACL18		KDD17	
	准确率	MCC	准确率	MCC
随机	50.102	0.002	49.850	-0.004
LSTM	53.876	0.101	52.679	0.027
Adv-ALSTM（Feng 等，2018）	57.200	0.148	53.050	0.052
DTML（Yoo 等，2021）	57.440	0.191	53.530	0.073
MAN-SF（Sawhney 等，2020年）	60.800	0.195	54.150	0.062
IDTLA（He、Siu 和 Si，2022）	58.200	0.168	—	—
STLAT-AT（Li、Dai 和 Zheng，2022）	60.610	0.224	55.730	0.1110
STLAT-TL（Li、Dai 和 Zheng，2022）	58.780	0.170	54.570	0.0864
STLAT（Li、Dai 和 Zheng，2022）	**64.560**	**0.297**	**58.270**	**0.159**
STST-MLP-T	56.624	0.126	52.860	0.045
STST-T	59.478	0.186	53.357	0.057
STST-MLP	62.509	0.249	54.319	0.075
STST	63.707	0.268	56.879	0.128

表3：STST 与其他基线方法和最先进的方法的结果比较。

此外，如果没有预测到任何股票会上涨，智能体就会暂缓交易，等待次日。测试期结束时，我们会统计累计收益和年化收益。随后，将两个数据集上的结果与标准普尔500指数的表现进行比较，以评估模拟智能体的整体表现。

结果

我们针对STST模型的大量候选超参数进行了网格搜索，见表1。我们发现效果最佳的超参数见表4。两个数据集的大多数参数都相同，只有少数参数例外。首先不同的是学习率。我们发现，较小的学习率 $2.5e^{-6}$能让模型在KDD数据集上的表现更好。其次不同的是编码器数量。对于ACL18数据集，STST使用四个编码器层时效果最佳；而对于KDD17数据集，使用八个编码器时效果最佳。最后，在训练过程中我们发现的另一处差异是LSTM层的数量。我们在ACL18中使用了两层堆叠LSTM层，而在KDD17中则使用了三层。这些差异可能是由数据集的规模造成的。KDD17数据集涵盖的时间跨度显著更长。由于市场会随时间显著变化，这给机器学习模型带来了难度更大的建模任务，也使模型更难进行泛化。

参数	ACL18	KDD17
学习率	$5.35e^{-6}$	$2.5e^{-6}$
批大小	32	32
预热步数	25000	25000
上下文窗口大小	32	32
编码器数量	4	8
注意力头数	4	4
模型维度	64	64
d_ff	2048	2048
前馈网络丢弃率	0.3	0.3
注意力丢弃率	0.1	0.1
LSTM层数	2	3
LSTM隐藏层大小	256	256
归一化类型	幂	幂
归一化位置	post	post

表4：STST模型的超参数选择。

模型性能

STST 方法在两个数据集上均表现良好。对于 ACL18数据集，其总体准确率和马修斯相关系数得分分别为 63.707% 和 0.268；对于 KDD17数据集，分别为 56.879% 和 0.128。该方法采用时空嵌入和长短期记忆网络分类器。STST 与其他基线模型见表3。我们的模型无需外部特征，也不必利用多个资产相关性，其结果便优于此前最先进方法中的所有方法，只有一种方法例外。遗憾的是，其整体表现略逊于完整的 STLAT 模型（Li、Dai 和 Zheng，2022）：该模型在 ACL18数据集上的准确率和 MCC 分别为 64.560% 和 0.297，在 KDD17数据集上分别为 58.270% 和 0.159。虽然表现不及完整的 STLAT 方法，但我们的 STST 方法显著优于未采用迁移学习的 STLAT 变体（STLAT-TL），在两个数据集上的准确率分别提升了 5.11% 和 6.84%；也显著优于未采用对抗训练的变体（STLAT-AD），在 ACL18数据集和 KDD17数据集上的准确率分别提升了 2.10% 和 4.26%。

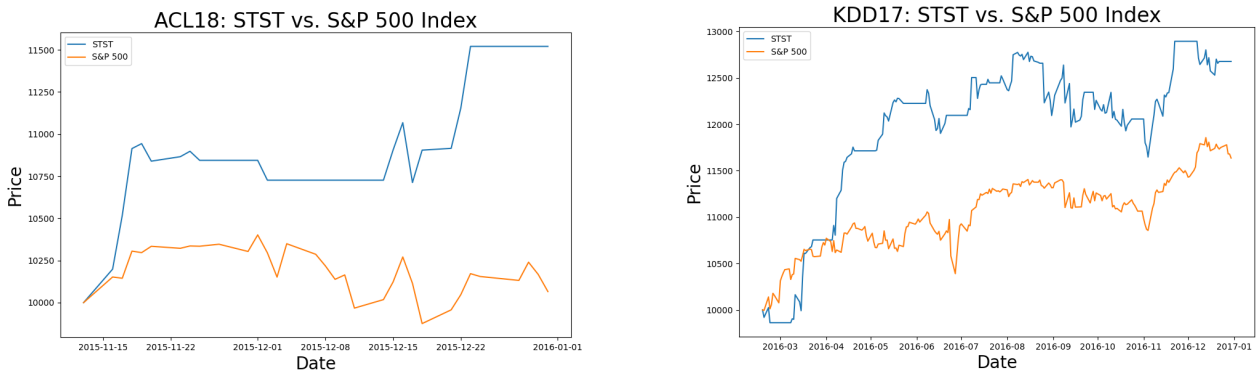


图4：ACL18（左）和KDD17数据集（右）的投资模拟累计利润与标准普尔500指数的对比

消融研究

在消融研究中，我们测试了 STST模型的三个变体，以展示时空注意力和 LSTM分类器对于最终模型的必要性和有效性。我们测试的第一个变体作为基线，采用了具有传统注意力的 Transformer 和 MLP分类器。在 ACL18数据集和 KDD17数据集上，该变体的准确率分别为 56.624% 和 52.860%。下一个变体是仅采用时间注意力和 LSTM-MLP分类器的 STST（STST-T）。加入长短期记忆网络后，该变体的表现显著优于基线，两个数据集上的准确率分别达到 59.478% 和 53.357%，比基线 STST模型最高提升 5.0%。最后，我们测试了采用时空嵌入和 MLP分类器的变体（STST-MLP）。该模型在两个数据集上的准确率分别为 62.509% 和 54.319%，比基线模型最高提升 10.4%。对于 ACL18数据集和 KDD17数据集，同时采用时空嵌入和 LSTM-MLP分类器的完整模型相较于基线，表现分别提升了 12.51% 和 6.64%。

投资模拟

为了检验模型在交易中的整体实际适用性，我们进行了投资模拟。在每个时间步，选择模型预测为正向的前5项资产来使用STST模型。这用于重新平衡股票投资组合，以最大化利润。图4展示了STST与标准普尔500指数的累计利润曲线。对于ACL18测试区间，STST的收益率为15.21%，而标准普尔500指数的收益率为0.66%。由于ACL18测试数据集的时间跨度较短，这相当于199%的年化收益率。在KDD17数据集上，STST的收益率为26.78%，而标准普尔500指数的收益率为16.36%，年化收益率因此达到31.24%。在这两个测试数据集上，我们的STST模型获得的收益率至少比标准普尔500指数高出10.41%；标准普尔500指数是评估交易策略时最常用的基准之一。分析各数据集的图表可以清楚地看到，在STST模型策略选择不进行交易的大多数情况下，也就是说，模型没有预测出任何股价上涨的股票时，持有标准普尔500指数反而会亏损。这表明STST模型能够充分预测股价下跌的时段。

讨论

总体而言，与以往及当前最先进的方法相比，我们的模型表现非常出色。STST的表现优于此前所有最先进的方法，与当前最先进的方法STLAT相比，在ACL18和KDD17数据集上的差距分别仅为0.853%和1.391%。

需要指出的是，作者提到，STLAT方法的重点在于训练过程，而非模型选择。因此，该方法主要关注用于微调的选择性采样过程，以及对抗训练的应用，而不是确定最优基础模型。在他们的研究中，STLAT的基础模型与Adv-ALSTM相同，而我们的模型表现显著优于Adv-ALSTM。Li等人提供了其方法在不使用迁移学习（STLAT-TL）以及不使用对抗训练（STLAT-AT）时的结果；我们的模型在ACL18和KDD17上的表现均优于这两种方法，优势至少分别达到3.097%和1.149%。此外，我们的模型还优于以往最先进的方法，包括Adv-ALSTM、DTML和MAN-SF，这些方法都更加侧重于模型选择。因此，尽管STLAT的训练方法总体上可能取得更好的结果，STST为股票走势预测提供了更好的基础。遗憾的是，STLAT方法尚未公开，因此我们无法与其进行更多比较。

在比较基线模型之后，我们进行了消融研究，以确定时空嵌入和 LSTM分类器对模型整体性能的总体影响。根据我们的结果，我们发现时空嵌入带来了最高 10.4% 的提升，而 LSTM分类器带来了

最高 5.0% 的提升，均优于基线 STST模型。由此可见，时空嵌入为 STST 方法带来的整体收益更大。不过，尽管时空嵌入带来的性能提升更为显著，我们也不能忽视 LSTM-MLP分类器的必要性。两者结合后，性能较基线大幅提升 12.5%。因此，我们可以得出结论：时空嵌入和 LSTM 分类器对于 STST 方法取得最佳结果都必不可少。

值得讨论的一个重要方面是 ACL18 和 KDD17 数据集所得准确率之间的差异，这一点可以从表3中我们模型和其他基线方法的准确率及 MCC 得分中看出。这很可能是由各数据集所涵盖的时间跨度不同造成的。KDD17数据集涵盖了 10 年，从 2007 年到 2017 年；而 ACL18数据集仅涵盖两年，从 2014 年到 2016 年。纵观历史，股票市场始终在不断演变，原因包括宏观经济、外交政策、技术进步、市场参与者等方面的变化。当观察更长的时间跨度时，就会发现市场如何变化，以及某些特定模式或评估技术可能不再像过去那样有效（Plastun 等，2021）。这很可能是两个数据集性能差异的根源。对于 ACL18数据集而言，其日期范围较短，可能不足以充分反映股票市场的历史，从而无法观察到不同类型的市场行为。此外，ACL18数据集仅包含 2014 年至 2016 年间的股票，而这一时期整体以牛市为主，市场动荡较少。KDD17数据集则涵盖了更长的时间段，始于 2007 年。因此，该数据集包含了 2008—2009 年的“大衰退”，这一事件对之后数年的市场产生了巨大影响。因此，KDD17数据集很可能涵盖了许多呈现不同类型行为的时期。对于机器学习模型而言，要在如此多样的样本上实现泛化，很可能困难得多。

未来

为了进一步提升 STST 模型的性能，我们计划在未来从多个不同方向开展研究。未来研究的首要方向是采用对抗训练。许多以往最先进的方法在将对抗训练用于股票走势预测时都取得了显著成效（Feng 等，2018）（Li、Dai 和 Zheng，2022）。通过添加随机扰动来进行对抗训练，有助于提升模型的泛化能力，并克服股票市场嘈杂且随机的特性。

我们希望进一步拓展研究的另一个领域，是利用相关资产和股票指数来更全面地了解市场，从而做出更准确的价格预测。尽管每只股票都有一系列可能影响其价格的因素，但许多宏观经济因素或政治事件往往会影响某个行业乃至整个市场的多只股票。不过，某些股票对这类变化的反应可能因多种因素而有所不同，有些价格变动也可能滞后。我们可以利用这一点来预测股票的走势。围绕这一理念进行交易通常称为“套利”。我们可以借助 STST，利用时空 Transformer 对同一行业中的多只股票进行编码，从而利用这一特点。随后，可以将每只股票的输出拼接起来，形成对市场的多重语境视图，方式类似于 DTML 的方法（ Yoo 等，2021 ）。接着，可以将这些结果输入长短期记忆网络或其他 Transformer，以关联各种模式并作出最终预测。

结论

由于股票市场蕴藏着巨大的获利潜力，大量研究致力于预测股票市场的各个方面，包括价格走势和交易信号。虽然许多交易者尝试手动完成这项工作，但由于预测时需要考量的数据量庞大，人类很快便难以准确完成这一任务。面对这些困难，再加上机器学习的兴起，研究者开始大量探索将机器学习应用于股票市场领域。不幸的是，由于市场具有随机性，股票数据中蕴含的关系十分复杂，往往很难提取。为此，我们提出 STST：一种将时空 Transformer 编码器与长短期记忆网络相结合、用于从股票数据中提取复杂关系的模型。我们展示了该模型在股价走势预测任务中的能力。STST 的表现优于此前四种最先进的方法，但仍略逊于当前最先进水平，在 ACL18 和 KDD17数据集上的差距分别为 0.853% 和 1.391%。尽管其表现未能超越 STLAT，我们证明了基础模型的学习能力强于此前用于股票市场预测的模型。STST 为未来研究奠定了坚实基础，且存在诸多可行的发展方向，有望使模型轻松超越当前最先进水平。除了模型的性能外，我们还通过模拟展示了模型在股票市场中的实际应用价值，其收益显著高于标普 500 市场指数。

参考文献

Abarbanell, J. S.; and Bushee, B. J. 1997. 基本面分析、未来收益与股价。《会计研究》, 35(1): 1–24. Chen, T.; and Guestrin, C. 2016. Xgboost：一种可扩展的树提升系统。载于第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集, 785–794. Dang, W.; Zhou, B.; Wei, L.; Zhang, W.; Yang, Z.; and Hu, S. 2021. *TS-Bert*：通过预训练模型*BERT*进行时间序列异常检测, 209–223. 国际标准书号 978-3-030-77963-4. Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT：用于语言理解的深度双向Transformer预训练。载于*NAACL*。

Dunne, M. 2015. 股票市场预测。科克大学学院。Edwards, R. D.; Magee, J.; and Bassetti, W. C. 2018. 股票趋势技术分析。CRC出版社。Feng, F.; Chen, H.; He, X.; Ding, J.; Sun, M.; and Chua, T.-S. 2018. 通过对抗训练增强股票走势预测。*arXiv*预印本 *arXiv:1810.09936*。Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM：一种用于点击率预测的基于因子分解机的神经网络。IJCAI 17, 1725, 1731. AAAI出版社。国际标准书号 9780999241103. He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. 用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集, 770–778. He, Q.; Siu, S.; and Si, Y. W. 2022. 基于实例的注意力深度迁移学习，用于股票走势预测。《应用智能》。Hearst, M.; Dumais, S.; Osuna, E.; Platt, J.; and Scholkopf, B. 1998. 支持向量机。《*IEEE*智能系统及其应用》, 13(4): 18–28. Hochreiter, S.; and Schmidhuber, J. 1997. 长短期记忆。《神经计算》, 9(8): 1735–1780. Kazemi, S. M.; Goel, R.; Eghbali, S.; Ramanan, J.; Sahota, J.; Thakur, S.; Wu, S.; Smyth, C.; Poupart, P.; and Brubaker, M. A. 2019. Time2Vec：学习时间的向量表示。*arXiv*, abs/1907.05321. Li, Y.; Dai, H.-N.; and Zheng, Z. 2022. 结合对抗训练的选择性迁移学习，用于股票走势预测。《连接科学》, 34(1): 492–510. Linsley, D.; Ki m, J.; Veerabadran, V.; Windolf, C.; and Serre, T. 2018. 使用水平门控循环单元学习长程空间依赖关系。载于 Bengio, S.; Wallach, H.; Larochelle, H.; Grau man, K.; Cesa-Bianchi, N.; and Garnett, R., 编，神经信息处理系统进展, 第31卷。Curran Associates, Inc. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*。Maberly, E. D. 1995. 尤里卡！尤里卡！星期一效应的发现应归功于古代抄写员。《金融分析师杂志》, 51(5): 10. Nair, V.; and Hinton, G. E. 2010. 整流线性单元改善受限玻尔兹曼机。载于*ICML*。Niaki, S.; and Hoseinzade, S. 2013. 使用人工神经网络和实验设计预测标准普尔500指数。《国际工业工程杂志》, 9: 1–9. Plastun, A.; Sibande, X.; Gupta, R.; and Wohar, M. E. 2021. 美国股票市场单日异常收益后价格效应的演变。《北美经济与金融杂志》, 57: 101405. Popel, M.; and Bojar, O. 2018. Transformer模型的训练技巧。*arXiv*预印本 *arXiv:1804.00247*。Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; and Cottrell, G. 2017. 一种用于时间序列预测的双阶段注意力循环神经网络。*arXiv*预印本 *arXiv:1704.02971*. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; 等. 2019. 语言模型是无监督多任务学习器。*OpenAI*博客, 1(8): 9. Sawhney, R.; Agarwal, S.; Wadhwa, A.; and Shah, R. R. 2020. 基于社交媒体文本和公司相关性的股票走势深度注意力预测。收录于*2020*年计算语言学实证方法会议论文集（*EMNLP*）, 8415–8426. 线上出版：计算语言学协会。Sen, J.; and Chaudhuri, T. 2017. 一种用于股价预测的稳健预测模型。收录于第*5*届商业分析与智能国际会议（*ICBAI 2017*）论文集，印度管理学院班加罗尔分校，印度。Shen, S.; Yao, Z.; Gholami, A.; Mahoney, M.; and Keutzer, K. 2020. Powernorm：重新思考Transformer中的批归一化。收录于国际机器学习会议, 8741–8751. PMLR. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. 注意力机制就是你所需要的一切。神经信息处理系统进展, 30. Wu, X.; Chen, H.; Wang, J.; Troiano, L.; Loia, V.; and Fujita, H. 2020. 基于深度强化学习方法的自适应股票交易策略。*InformationSciences*, 538: 142 – 158. Xiong, R.; Yang, Y.; He, D.; Zheng, K.; Zheng, S.; Xing, C.; Zhang, H.; Lan, Y.; Wang, L.; and Liu, T.-Y. 2020. 关于Transformer架构中的层归一化。收录于第*37*届国际机器学习会议论文集, ICML'20. JMLR.org. Xu, Y.; and Cohen, S. B. 2018. 基于推文和历史价格的股票走势预测。收录于*ACL*。Yoo, J.; Soun, Y.; Park, Y.-c.; a nd Kang, U. 2021. 通过带有多层级上下文的数据轴Transformer进行准确的多变量股票走势预测。KDD 21, 2037, 2045. 纽约，NY，美国：计算机协会。ISBN 9781450383325. Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. 一种用于多变量时间序列表示学习的基于Transformer的框架。纽约，NY，美国：计算机协会。ISBN 9781450383325. Zhang, L.; Aggarwal, C.; and Qi, G.-J. 2017. 通过发现多频交易模式进行股价预测。收录于第*23*届*ACM*知识发现与数据挖掘国际会议论文集, KDD '17, 2141, 2149. 纽约，NY，美国：计算机协会。ISBN 9781450348874. Zhang, X.; Liu, S.; and Zheng, X. 2021. 基于深度分解机和注意力机制的股价走势预测。*Mathematics*, 9(8)。