姓名手稿编号
(将由编辑插入)

多元时间序列预测的时间模式关注

石顺尧* • 孙凡锫* • 李鸿祎

收稿日期:date /收稿日期:date

摘要多元时间序列数据的预测，如电力消耗、太阳能发电和复调钢琴曲的预测，具有许多有价值的应用。然而，时间步长和序列之间复杂的非线性相互依赖关系使这项任务复杂化。为了获得准确的预测，对时间序列数据中的长期依赖性进行建模是至关重要的，这可以通过具有注意机制的递归神经网络(RNNs)来实现。典型的注意机制会在之前的每个时间步检查信息，并选择相关信息来帮助生成输出;然而，它无法捕捉跨多个时间步长的时间模式。在本文中，我们建议使用一组滤波器来提取时不变的时间模式，类似于将时间序列数据转换为其“频域”。然后，我们提出了一种新颖的注意力机制来选择相关的时间序列，并利用其频域信息进行多元预测。我们将提出的模型应用于几个现实世界的任务，并在几乎所有情况下实现了最先进的性能。我们的源代码可在 https://github.com/gantheory/TPA-LSTM 获得。

*表示同等贡献。

本工作得到了台湾科技部的财政支持。

石顺尧

国立台湾大学电子邮件:
shunyaoshih@gmail.com 孙凡锫

国立台湾大学电子邮
件:b03901056@ntu.edu.tw李鸿
祎

国立台湾大学电子邮
件:hungyilee@ntu.edu.tw

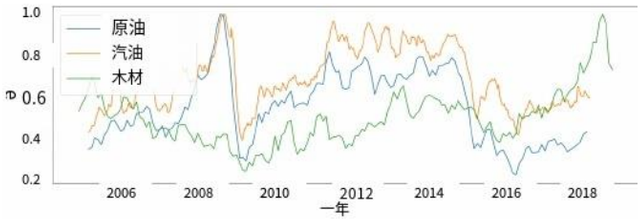

图 1 原油、汽油和木材的历史价格。为简单起见，单位省略，尺度归一化。

1 介绍

在日常生活中，时间序列数据无处不在。我们观察由传感器在离散时间步长上生成的演化变量，并将其组织成时间序列数据。例如，家庭用电量、道路占用率、货币汇率、太阳能发电量，甚至音符都可以被视为时间序列数据。在大多数情况下，收集的数据通常是多变量时间序列(MTS)数据，例如由当地电力公司跟踪的多个客户的用电量。不同序列之间可能存在复杂的动态相互依存关系，这些相互依存关系很重要，但难以捕获和分析。

分析师往往寻求根据历史数据预测未来。不同序列之间的相互依赖关系建模得越好，预测就越准确。例如，如图 1[1]所示，原油价格严重影响汽油价格，但对木材价格的影响较小。因此，在认识到汽油是由原油生产而木材不是由原油生产的情况下，我们可以用原油价格来预测汽油的价格。

在机器学习中，我们希望模型能够从数据中自动学习到这种相互依赖关系。机器学习已经被应用于时间序列分析，用于分类和预测[G。张和胡(1998)，张(2003)，Lai 等(2018)Lai，张，杨，刘，秦等(2017)秦，Song, Cheng, Cheng, Jiang, and Cottrell。在分类中，机器学习为时间序列分配标签，例如通过读取医疗传感器的值来评估患者的诊断类别。在预测方面，机器根据过去观察到的数据预测未来的时间序列。例如，可以根据历史测量结果预测未来几天、几周或几个月的降水。我们越想提前预测，就越难预测。

当涉及到使用深度学习的 MTS 预测时，经常使用循环神经网络(rnn)[David E. Rumelhart and Williams(1986)， J.Werbos(1990)， Elman(1990)]。然而，在时间序列分析中使用 rnn 的一个缺点是它们在管理长期依赖关系方面的弱点，

1 来源:https://www.eia.gov 和 https://www.investing.com

例如，每日记录序列中的年模式[Kyunghyun Cho 和 Bengio(2014)]。注意力机制[Luong 等人(2015)Luong、Pham 和 Manning, Bahdanau 等人(2015)Bahdanau、Cho 和 Bengio]最初用于编码器-解码器[Sutskever 等人(2014)Sutskever、Vinyals 和 Le]网络，在一定程度上缓解了这一问题，从而提高了 RNN [Lai 等人(2018)Lai、Chang、Yang 和 Liu]的有效性。

在本文中，我们提出了一种新的 MTS 预测注意机制——时间模式注意，其中我们使用“时间模式”一词来指代跨多个时间步长的任何时不变模式。典型的注意机制识别与预测相关的时间步长，并从这些时间步长中提取信息，这对 MTS 预测有明显的局限性。考虑图 1 中的例子。为了预测汽油的价值，机器必须学会关注“原油”而忽略“木材”。在我们的时间模式注意中，机器不是像典型的注意机制那样选择相关的时间步长，而是学习选择相关的时间序列。

此外，时间序列数据通常需要明显的周期性时间模式，这对预测至关重要。然而，跨越多个时间步长的周期性模式对于典型的注意力机制来说很难识别，因为它通常只关注几个时间步。在时间模式注意中，我们引入卷积神经网络(CNN) [LeCun and Bengio(1995)， a . Krizhevsky and Hinton(2012)]从每个个体变量中提取时间模式信息。

本文的主要贡献总结如下:

-我们引入了一个新的注意力概念，在这个概念中，我们选择相关的变量，而不是相关的时间步长。该方法简单、通用，适用于 RNN。

-我们使用玩具示例来验证我们的注意力机制使模型能够提取时间模式，并关注不同时间序列的不同时间步长。

-通过从周期和部分线性到非周期和非线性任务的真实世界数据的实验结果证明，我们表明所提出的注意力机制在多个数据集上实现了最先进的结果。

-在我们的注意力机制中学习的 CNN 过滤器表现出有趣和可解释的行为。

本文的其余部分组织如下。在第 2 节中，我们回顾了相关的工作，在第 3 节中，我们描述了背景知识。然后，在第 4 节中，我们描述了提出的注意力机制。接下来，我们在第 5 节中介绍并分析了我们对玩具示例的注意机制，在第 6 节中介绍并分析了 MTS 和复调音乐数据集的注意机制。最后，我们在第 7 节中进行总结。

2 相关工作

线性单变量时间序列预测最著名的模型是自回归积分移动平均(ARIMA)[G]。E. Box和Ljung(2015)]，它包含了其他自回归时间序列模型，包括自回归(AR)、移动平均(MA)和自回归移动平均(ARMA)。此外，线性支持向量回归(SVR) [Cao and Tay(2003)， Kim(2003)]将预测问题视为具有时变参数的典型回归问题。然而，这些模型大多局限于线性单变量时间序列，不能很好地扩展到MTS。为了预测MTS数据，提出了向量自回归(VAR)，一种基于ar的模型的推广。VAR可能是MTS预测中最著名的模型。然而，无论是基于ar的还是基于var的模型都不能捕捉到非线性。出于这个原因，大量的努力已经投入到基于核方法的非线性时间序列预测模型[Chen等人(2008)]、Chen、Wang和Harris]、集合[Bouchachia和Bouchachia(2008)]、高斯过程[Frigola和Rasmussen(2014)]或状态切换[Tong和Lim(2009)]。然而，这些方法应用预定的非线性，可能无法识别不同形式的非线性对不同的MTS。

最近，深度神经网络由于其捕获非线性相互依赖关系的能力而受到了大量关注。长短期记忆(LSTM) [Hochreiter和Schmidhuber(1997)]是递归神经网络的一种变体，在几个NLP任务中显示出了令人满意的结果，也被用于MTS预测。该领域的工作始于使用朴素RNN [J]。Connor和Martin(1991)]，使用结合了ARIMA和多层感知器的混合模型进行改进[j]。Zhang和Hu(1998)， Zhang(2003)， Jain和Kumar(2007)]，然后最近进展到使用RNN的动态玻尔兹曼机[Dasgupta和Os- ogami(2017)]。虽然这些模型可以应用于MTS，但它们主要针对单变量或双变量时间序列。

据我们所知，长期和短期时间序列网络(LSTNet) [Lai等人(2018)Lai, Chang, Yang, and Liu]是第一个专门为多达数百个时间序列的MTS预测设计的模型。LSTNet使用cnn捕获短期模式，LSTM或GRU用于记忆相对长期的模式。然而，在实践中，由于训练不稳定性和梯度消失问题，LSTM和GRU不能记住非常长期的相互依赖关系。为了解决这个问题，LSTNet增加了一个循环跳过层或一个典型的注意机制。整体模型的另一部分是传统的自回归，这有助于减轻神经网络的尺度不敏感性。尽管如此，与我们提出的注意机制相比，LSTNet有三个主要缺点:(1)必须手动调整循环跳过层的跳过长度以匹配数据的周期，而我们提出的方法可以自己学习周期模式;(2) LSTNet-Skip模型是专门为具有周期性模式的MTS数据设计的，而我们提出的模型从实验中可以看出，它简单且适用于各种数据集，甚至是非周期性的数据集

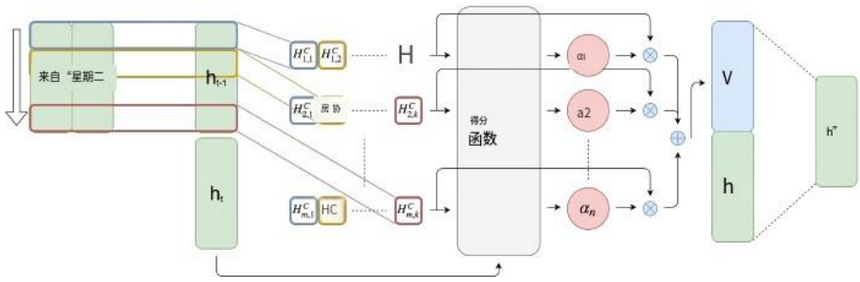

图 2 提出的注意机制。ht 表示 RNN 在时间步长 t 时的隐藏状态。有 k 个长度为 w 的 1-D CNN 滤波器，用不同颜色的矩形表示。然后，每个滤波器对 m 个隐藏状态的特征进行卷积，并产生一个 m 行 k 列的矩阵 HC。接下来，评分函数通过比较当前隐藏状态 ht 来计算每行 HC 的权重。最后，对权重进行归一化，将 HC 的行按其对应的权重进行加权求和，生成 Vt，最后将 Vt 与 ht 进行连接，并进行矩阵乘法，生成用于生成最终预测值的 hott。

的人;(3) LSTNet-Attn 模型中的注意层与典型的注意机制一样选择相关的隐藏状态，而我们提出的注意机制选择相关的时间序列，这是一种更适合 MTS 数据的机制。

3 预赛

在本节中，我们简要描述了与我们提出的模型相关的两个基本模块:RNN 模块和典型的注意力机制。

3.1 递归神经网络

给定一个信息序列{$x_1, x_2, \cdots, x_t$}，其中 $x_i \in R^n$, RNN 一般定义一个循环函数 F，并计算每个时间步长 t 的 $h_t \in {}^m R$ 为

$$h_t = F(h_{t-1}, x_t) \tag{1}$$

函数 F 的实现取决于使用哪种 RNN 单元。

长短期记忆(Long - short memory, LSTM) [Hochreiter and Schmidhuber(1997)]细胞被广泛使用，它们的循环功能略有不同:

$$h_t, c_t = F(h_{t-1}, c_{t-1}, x_t), \tag{2}$$

由以下方程定义:

$$i_t = sigmoid(W_{x_i} x_t + W_{h_i} h_{t-1}) \quad (4)$$

$$f_t = sigmoid(W_{x_f} x_t + W_{h_f} h_{t-1}) \quad (5)$$

$$o_t = sigmoid(W_{x_o} x_t + W_{h_o} h_{t-1}) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{x_g} x_t + W_{h_g} h_{t-1}) \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (6)$$

其中 it, ft, ot, ct∈Rm,Wxi,Wxf,Wxo,Wxg∈Rm×n,Whi, Whf,

W_{h_o} and $W_{h_g} \in \mathbb{R}^{m \times m}$, and

denotes element-wise multiplication.

3.2 典型的注意机制

在 RNN 中的典型注意机制[Luong 等人(2015)， Luong, Pham, and Manning, Bahdanau 等人(2015)， Bahdanau, Cho 和 Bengio]中，给定先前的状态 $H_1 = \{H_2, H, \cdots, h_{t-1}\}$，从之前的状态中提取的上下文向量 vis_t。vis_t 每列 hin_i H 的加权和，表示与当前时间步长相关的信息。Vis_t 进一步与当前状态 h 结合，t 得到预测结果。

假设一个评分函数 $f: R^m \times R^m 7 \rightarrow R$，它计算其输入向量之间的相关性。形式上，我们有以下公式来计算上下文向量 v_t:

$$\alpha_i = \frac{\exp(f(h_i, h_t))}{\sum_{j=1}^{t-1} \exp(f(h_j, h_t))} \quad (8)$$

$$v_t = \sum_{i=1}^{t-1} \alpha_i h_i. \quad (8)$$

4 时间模式注意

虽然以前的工作主要集中在通过不同的设置来改变基于注意力的模型的网络架构，以提高在各种任务上的性能，但我们认为在 RNN 上应用典型的注意力机制进行 MTS 预测存在一个关键缺陷。典型的注意力机制选择与当前时间步相关的信息，上下文向量 t 是之前 RNN 隐藏状态列向量的加权和，$H = \{H_1, 2H, \cdots h\}$。t-1 这种设计适合于每个时间步长只包含一条信息的任务，例如，每个时间步长对应一个单词的 NLP 任务。如果每个时间步长中有多个变量，它就无法忽略在预测效用方面有噪声的变量。此外，由于典型的注意力机制在多个时间步长中平均信息，因此它无法检测到对预测有用的时间模式。

所建议模型的概述如图 2 所示。在所提出的方法中，给定之前的 RNN 隐藏状态 H∈rmx(t− 1)，所提出的注意机制基本上关注其行向量。行上的注意权值选择那些有助于预测的变量。由于上下文向量现在是 t 包含多个时间步长的信息的行向量的加权和，因此它捕获了时间信息。

4.1 问题表述

在 MTS 预测中，给定一个 MTS, X1= {X2, X，…， xt− 1}，其中 xi ∈n r 表示时刻 i 的观测值，任务是预测 xt− 1+Δ 的值，其中Δ 是相对于不同任务的固定视界。我们将相应的预测记为 yt− 1+Δ ，将真值记为 y′ t− 1+Δ =xt− 1+Δ 。此外，对于每个任务，我们只使用{xt -w, xt -w+1，…。， xt− 1}来预测 x，t− 1+Δ 其中 w 是窗口大小。这是一种常见的做法[Lai 等人(2018)Lai, Chang, Yang, and Liu, Qin 等人(2017)Qin, Song, Cheng, Cheng, Jiang, and Cottrell]，因为假设窗口前没有有用的信息，因此输入是固定的。

4.2 使用 CNN 进行时间模式检测

CNN 的成功在很大程度上取决于它捕捉各种重要信号模式的能力;因此，我们使用 CNN 对 h 的行向量应用 CNN 滤波器来增强模型的学习能力。具体来说，我们有 k 个滤波器 Ci ∈R1×T，其中 T 是我们关注的最大长度。如果未指定，我们假设 T = w。卷积运算产生 HC ∈n×k r，其中 i,jC H 表示第 i 行向量和第 j 个滤波器的卷积值。形式上，这个运算由

$$H_{i,j}^{C}=\sum_{l=1}^{w}H_{i,(t-w-1+l)}\times C_{j,T-w+l}. \tag{10}$$

4.3 建议的注意机制

我们计算 vast 为 H 的行向量的加权和。C 下面定义的是评价 k 相关性 m 的评分函数 f: rx R7→R:

$$f(H_i^C,h_t)=(H_i^C)^\top W_a h_t, \tag{11}$$

式中，HiC 为 hc 的第 i 行，Wa∈Rk×m。注意权值 αi 得为

$$\alpha_i=sigmoid(f(H_i^C,h_t)). \tag{12}$$

请注意，我们使用的是 sigmoid 激活函数而不是 softmax，因为我们期望不止一个变量对预测有用。

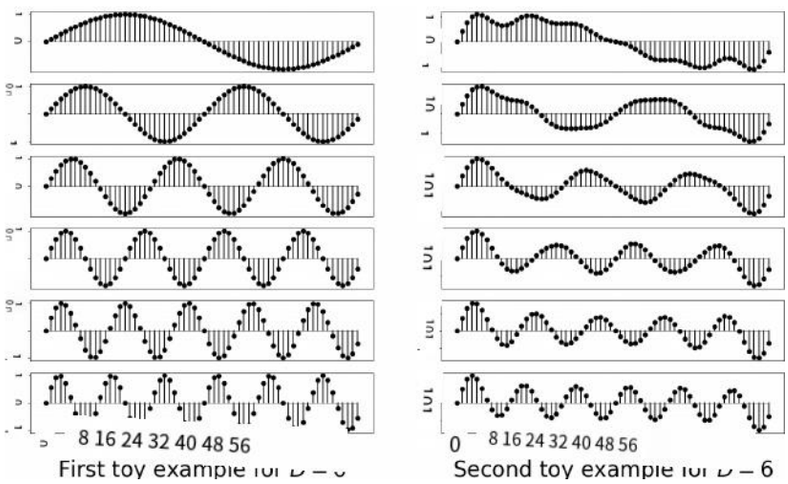

图 3 D = 6 时，第一类无相互依赖的玩具样例(左)和第二类有相互依赖的玩具样例(右)的可视化，这意味着每个样例中有 6 个时间序列。

完成这个过程，对 H^C 的行向量进行 α_i 加权，得到上下文向量 $v_t \in R^k$，

$$v_t = \sum_{i=1}^{n} \alpha_i H_i^C. \tag{13}$$

然后对 v_t and h 进行积分 $_t$，得到最终的预测结果

$$h'_t = W_h h_t + W_v v_t, \tag{14}$$

$$y_{t-1+\Delta} = W_{h'} h'_t, \tag{15}$$

where $h_t, h'_t \in \mathbb{R}^m$, $W_h \in \mathbb{R}^{m \times m}$, $W_v \in \mathbb{R}^{m \times k}$, and $W_{h'} \in \mathbb{R}^{n \times m}$ and $y_{t-1+\Delta} \in$ Rn 。

5 对玩具示例的建议关注分析

为了阐述传统注意机制的失效和相互依赖的影响，我们研究了不同注意机制在两类人工构建的玩具样例上的表现。

在第一类玩具样例中，将第 i 个时间序列的第 t 个时间步长定义为 $\sin(2\pi it64)$，即每个时间序列都是具有不同周期的正弦波。注意，在第一类中，任意两个时间序列都是相互独立的，所以不存在相互依赖。

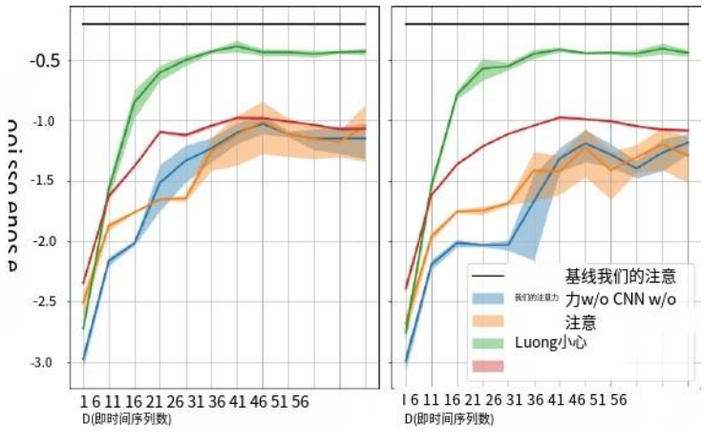

图 4 第一类没有相互依赖关系的玩具样本(左)和第二类具有相互依赖关系的玩具样本(右)的平均绝对损失和 log10 的标准差范围，两者都在 10 次运行中。基线表示所有预测值为零时的损失。

第二类玩具样例通过混合时间序列来增加第一类的相互依赖性，因此第 i 类时间序列的第 t 个时间步长公式为:

$$sin(\frac{2\pi it}{64}) + \frac{1}{D-1}\sum_{j=1,j\neq i}^{D} sin(\frac{2\pi jt}{64}), \tag{16}$$

式中 D 为时间序列的个数。对于 D = 6，这两种类型的玩具示例都在图 3 中可视化。

以下分析中的所有模型都使用窗口大小 w = 64，水平Δ = 1 和相似数量的参数进行训练。在这个设置中，我们的每个玩具示例由 64 个样本组成。第一个样本中的每个时间序列包含从 t = 0 到 63 的 Eq. 16 的值，我们可以移动一个时间步来获得值从 t = 1 到 64 的第二个样本。对于最后一个样本，我们使用 t = 63 到 126 之间的值作为相应的输入序列。注意，t = 64 到 127 的值等于 t = 0 到 63 的值。我们在 D ={1,6,11，⋯}的两种玩具样本上训练了 200 个 epoch 的模型。， 56}并记录训练中的平均绝对损失。没有验证和测试数据，因为本节的目的是证明我们的注意力比典型的注意力具有更大的能力来拟合 MTS 数据，而不是我们注意力的泛化性。结果如图 4 所示。

5.1 传统注意机制失效

直观地说，对于第一个玩具示例，模型可以通过记忆恰好在一个周期之前出现的值来准确预测下一个值。然而，我们知道不同的时间序列有不同的周期，这意味着

为了有一个好的预测，模型应该能够回顾不同序列的不同时间步长。从这一点来看，很明显，传统注意力机制的失败来自于只提取了一个之前的时间步长，而贴现了其他时间步长的信息。另一方面，我们的注意机制关注 CNN 滤波器从 RNN 隐藏状态的行向量中提取的特征，使模型能够跨多个时间步选择相关信息。

上述解释可以从图 4 的左图中得到验证，在图 4 中我们观察到，当 D ?1，与其他的相比。请注意，所有模型都有相似数量的参数，这意味着没有注意的 LSTM 与具有 Luong 注意的 LSTM 相比具有更大的隐藏大小。因此，当 D ?时，不加注意的 LSTM 优于 Luong 注意的 LSTM。1、因为较大的隐量有助于模型进行预测，而 Luong 几乎是无用的。相反，我们的注意力是有用的，所以有我们的注意力的 LSTM 比没有注意的 LSTM 平均要好，即使它的隐藏大小更小。此外，将 CNN 从我们的注意力中移除，会导致与表 4 中的“Sigmoid - W/o CNN”单元相同的模型，这并不影响性能，这意味着我们的特征关注是不可或缺的。

5.2 相互依赖关系的影响

当 MTS 数据中存在相互依赖关系时，利用相互依赖关系进一步提高预测精度是可取的。图 4 中的右图显示，具有 Luong 注意的 LSTM 和没有注意的 LSTM 都没有从增加的相互依赖性中受益，因为损失值保持不变。另一方面，当存在相互依赖关系时，所提出的注意力的 LSTM 损失较低，这表明我们的注意力成功地利用了相互依赖关系来促进 MTS 预测。同样，在这种情况下，将 CNN 从我们的注意力中移除并不影响性能。

6 实验与分析

在本节中，我们首先描述了我们进行实验的数据集。接下来，我们展示了我们的实验结果和针对 LSTNet 的预测的可视化。然后，我们讨论了烧蚀研究。最后，我们分析了 CNN 滤波器在何种意义上类似于 DFT 中的基。

6.1 数据集

为了评估所提出的注意机制的有效性和泛化能力，我们使用了两种不同类型的数据集:典型 MTS 数据集和复调音乐数据集。

典型的 MTS 数据集由[Lai 等人(2018)Lai, Chang, Yang, and Liu];有四个数据集:

-太阳能 [2]:2006 年阿拉巴马州光伏电站的太阳能发电量数据。

-交通 [3]:加州交通部提供的两年(2015-2016)数据，描述了旧金山湾区高速公路的道路占用率(介于 0 到 1 之间)。

—用 [4] 电量:记录 321 个客户的用电量，单位为千瓦时。

-汇率:1990 年至 2016 年 8 个国外国家(澳大利亚、英国、加拿大、中国、日本、新西兰、新加坡、瑞士)的汇率。

这些数据集是真实世界的数据，包含线性和非线性的相互依赖关系。此外，太阳能、交通和电力数据集表现出强烈的周期性模式，表明每天或每周的人类活动。根据 LSTNet 作者的说法，所有数据集中的每个时间序列都按时间顺序分为训练集(60%)、验证集(20%)和测试集(20%)。

相比之下，下面介绍的复调音乐数据集要复杂得多，从某种意义上说，没有明显的线性或重复模式存在:

- MuseData [Nicolas Boulanger-Lewandowski and Vincent(2012)]:以 MIDI 格式收录了各种古典音乐作曲家的音乐作品。
- LPD-5-Cleansed[董浩文和杨(2018)，拉斐尔(2016)]:21,425 首包含鼓、钢琴、吉他、贝斯和弦乐的多轨钢琴卷。

为了在这些数据集上训练模型，我们将每个演奏的音符视为 1，否则为 0(即音乐休息)，并将一个节拍设置为一个时间步长，如表 1 所示。给定由 16 拍组成的 4 小节的演奏音符，任务是预测下一个时间步的每个音高是否被演奏。对于训练、验证和测试集，我们遵循原始的 MuseData 分离，分为 524 个训练片段、135 个验证片段和 124 个测试片段。然而，在之前的研究中，lpd -5-cleaned 并没有被拆分[董浩文和杨(2018)，拉斐尔(2016)];因此我们随机地将其分成训练集(80%)、验证集(10%)和测试集(10%)。lpd -5- cleaned 数据集的大小比其他数据集大得多，因此我们决定使用较小的验证和测试集。

2 http://www.nrel.gov/grid/solar-power-data.html

3 http://pems.dot.ca.gov

4 https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

数据集	T	D	年代	B
太阳能	52560 年	137	10 分钟	172 米
交通	17544 年	862	1 小时	130 米
电	26304 年	321	1 小时	91 米
汇率	7588 年	8	1 天	534 K
MuseData	216 - 102552	128	1 击败	4.9 米
lpd -5 清洁	1072 - 1917952	128	1 击败	1.7 克

表 1 所有数据集的统计，其中 L 为时间序列的长度，D 为时间序列的个数，S 为采样间隔，B 为数据集的大小(以字节为单位)。MuseData 和 lpd -5 清洁都有不同长度的时间序列，因为音乐作品的长度是不同的。

典型 MTS 数据集和复调音乐数据集的主要区别在于，典型 MTS 数据集中的标量是连续的，而复调音乐数据集中的标量是离散的(0 或 1)。表 1 总结了典型 MTS 数据集和复调音乐数据集的统计数据。

6.2 比较的方法

在典型的 MTS 数据集上，我们将本文提出的模型与以下方法进行了比较:

- AR:标准自回归模型。

LRidge:具有 l2 正则化的 VAR 模型:最流行的 MTS 预测模型。

- LSVR:具有 SVR 目标函数的 VAR 模型[V]。Vapnik(1997)]。

- GP:高斯过程模型[Frigola-Alcade(2015)， S. Roberts and Aigrain(2011)]。

- SETAR:自激阈值自回归模型，经典的单变量非线性模型[Tong and Lim(2009)]。

—LSTNet- skip:具有循环跳过层的 LSTNet。

- LSTNet- attn:带注意层的 LSTNet。

AR、LRidge、LSVR、GP 和 SETAR 是传统的基线方法，而 LSTNet- skip 和 LSTNet-Attn 是基于深度神经网络的最先进方法。

然而，由于传统的基线方法和 LSTNet 由于其非线性和缺乏周期性而不适合复调音乐数据集，我们使用 LSTM 和 Luong 注意力的 LSTM 作为基线模型来评估所提出的模型在复调音乐数据集上的效果:

—LSTM:第 3 节介绍的 RNN 单元。

-具有 Luong 注意的 LSTM:具有注意机制评分函数的 LSTM，其 f(hi, ht)= (hi)>Wht，其中 W∈Rm×m [Luong 等人(2015)Luong, Pham, and Manning]。

6.3 模型设置和参数设置

对于所有实验，我们在 RNN 模型中使用 LSTM 单元，并将 CNN 滤波器的数量固定为 32。此外，受 LSTNet 的启发，我们在对典型 MTS 数据集进行训练和测试时，在我们的模型中加入了一个自回归组件。

对于典型的 MTS 数据集，我们使用 LSTNet 对可调参数进行网格搜索。具体来说，在太阳能、交通和电力方面，窗口大小 w 的取值范围为{24,48,96,120,144,168}，隐藏单元数 m 的取值范围为{25,45,70}，速率为 0.995 的指数学习率衰减步长取值范围为{200,300,500,1000}。在汇率中，这三个参数分别为{30,60}、{6,12}和{120,200}。两种类型的数据归一化也被视为网格搜索的一部分:一种是通过自身的最大值对每个时间序列进行归一化，另一种是通过整个数据集上的最大值对每个时间序列进行归一化。最后，我们使用绝对损失函数和 Adam^{-3}，太阳能、交通和电力的学习率为 10，汇率的学习率为 $3 \cdot 10^{-3}$。对于 AR、LRidge、LSVR 和 GP，我们遵循 LSTNet 论文[Lai et al.(2018)Lai, Chang, Yang, and Liu]中报告的参数设置。对于 SETAR，我们在{24,48,96,120,144,168}上搜索太阳能、交通和电力的嵌入维数，将汇率的嵌入维数固定为 30。我们的方法和 LSTNet 之间的两个不同设置是:(1)我们有两种数据规范化方法可供选择，而 LSTNet 只使用第一种类型的数据规范化;以及(2)窗口大小 w 上的网格搜索是不同的。

对于用于复调音乐数据集的模型，包括以下小节中的基线和建议模型，我们对所有 mn 使用 3 层，如[Chuan and Herremans(2018)]所做的那样，并通过调整 LSTM 单元的数量将可训练参数固定为 5.10^6 左右，以公平地比较不同的模型。此外，我们使用了具有 10^{-5} 学习率和交叉熵损失函数的 Adam 优化器。

6.4 评估指标

在典型的 MTS 数据集上，由于我们将所提出的模型与 LST- Net 进行了比较，我们遵循了相同的评估指标:RAE、RSE 和 CORR。第一个指标是相对绝对误差(RAE)，定义为

$$\mathrm{RAE} = \frac{\sum_{t=t_0}^{t_1} \sum_{i=1}^{n} |(y_{t,i} - y^3_{t,i})|}{\sum_{t=t_0}^{t_1} \sum_{i=1}^{n}} \quad 。 \tag{17}$$

RAE	太阳能				交通			
地平线	3	6	12	24	3	6	12	24
基	0.1846	0.3242	- 0.5637	0.9221	0.4491	0.4610	- 0.4700	0.4696
LRidge	0.1227	0.2098	- 0.4070	0.6977	0.4965	0.5115	- 0.5198	0.4846
LSVR	0.1082	0.2451	0.4362	0.6180	0.4629	0.5483	0.7454	0.4761
全科医生	0.1419	0.2189	0.4095	0.7599	0.5148	0.5759	0.5316	0.4829
伴随着	0.1285	0.1962	0.2611	0.3147	0.3226	0.3372	0.3368	0.3348
LSTNet-Skip	0.0985	0.1554	0.2018	0.3551	0.3287	0.3627	0.3518	0.3852
LSTNet-Attn	0.0900	0.1332	0.2202	0.4308	0.3196	0.3277	0.3557	0.3666
我们的模型	0.0918 ± 0.0005	0.1296 ± 0.0008	0.1902 ± 0.0021	0.2727 ± 0.0045	0.2901 ± 0.0095	0.2999 ± 0.0022	0.3112 ± 0.0015	0.3118 ± 0.0034

雷	电				汇率			
地平线	3.	6	12	24	3.	6	12	24
基于“增大化现实”技术	0.0579	0.0598	0.0603	0.0611	0.0181	0.0224	0.0291	0.0378
LRidge	0.0900	0.0933	0.1268	0.0779	0.0144	0.0225	0.0358	0.0602
LSVR	0.0858	0.0816	0.0762	0.0690	0.0148	0.0231	0.0360	0.0576
全科医生	0.0907	0.1137	0.1043	0.0776	0.0230	0.0239	0.0355	0.0547
伴随着	0.0475	0.0524	0.0545	0.0565	0.0136	0.0199	0.0288	0.0425
LSTNet-Skip	0.0509	0.0587	0.0598	0.0561	0.0180	0.0226	0.0296	0.0378
LSTNet-Attn	0.0515	0.0543	0.0561	0.0579	0.0229	0.0269	0.0384	0.0517
我们的模型	0.0463 ± 0.0007	0.0491 ± 0.0007	0.0541 ± 0.0006	0.0544 ± 0.0007	0.0139 ± 0.0001	0.0192 ± 0.0002	0.0280 ± 0.0006	0.0372 ± 0.0005

交易所	太阳能				交通			
地平线	3.	6	12	24	3.	6	12	24
基于“增大化现实”技术	0.2435	0.3790	0.5911	0.8699	0.5991	0.6218	0.6252	0.6293
LRidge	0.2019	0.2954	0.4832	0.7287	0.5833	0.5920	0.6148	0.6025
LSVR	0.2021	0.2999	0.4846	0.7300	0.5740	0.6580	0.7714	0.5909
全科医生	0.2259	0.3286	0.5200	0.7973	0.6082	0.6772	0.6406	0.5995
伴随着	0.2374	0.3381	0.4394	0.5271	0.4611	0.4805	0.4846	0.4898
LSTNet-Skip	0.1843	0.2559	0.3254	0.4643	0.4777	0.4893	0.4950	0.4973
LSTNet-Attn	0.1816	0.2538	0.3466	0.4403	0.4897	0.4973	0.5173	0.5300
我们的模型	0.1803 ± 0.0008	0.2347 ± 0.0017	0.3234 ± 0.0044	0.4389 ± 0.0084	0.4487 ± 0.0180	0.4658 ± 0.0053	0.4641 ± 0.0034	0.4765 ± 0.0068

交易所	电				汇率			
地平线	3.	6	12	24	3.	6	12	24
基于“增大化现实”技术	0.0995	0.1035	0.1050	0.1054	0.0228	0.0279	0.0353	0.0445

LRidge	0.1467	0.1419	0.2129	0.1280	0.0184	0.0274	0.0419	0.0675
LSVR	0.1523	0.1372	0.1333	0.1180	0.0189	0.0284	0.0425	0.0662
全科医生	0.1500	0.1907	0.1621	0.1273	0.0239	0.0272	0.0394	0.0580
伴随着	0.0901	0.1020	0.1048	0.1009	0.0178	0.0250	0.0352	0.0497
LSTNet-Skip	0.0864	0.0931	0.1007	0.1007	0.0226	0.0280	0.0356	0.0449
LSTNet-Attn	0.0868	0.0953	0.0984	0.1059	0.0276	0.0321	0.0448	0.0590
我们的模型	0.0823 ±0.0012	0.0916 ±0.0018	0.0964 ±0.0015	0.1006 ±0.0015	0.0174 ±0.0001	0.0241 ±0.0004	0.0341 ±0.0011	0.0444 ±0.0006

相关系数	太阳能				交通			
地平线	3.	6	12	24	3.	6	12	24
基于“增大化现实”技术	0.9710	0.9263	0.8107	0.5314	0.7752	0.7568	0.7544	0.7519
LRidge	0.9807	0.9568	0.8765	0.6803	0.8038	0.8051	0.7879	0.7862
LSVR	0.9807	0.9562	0.8764	0.6789	0.7993	0.7267	0.6711	0.7850
全科医生	0.9751	0.9448	0.8518	0.5971	0.7831	0.7406	0.7671	0.7909
伴随着	0.9744	0.9436	0.8974	0.8420	0.8641	0.8506	0.8465	0.8443
LSTNet-Skip	0.9843	0.9690	0.9467	0.8870	0.8721	0.8690	0.8614	0.8588
LSTNet-Attn	0.9848	0.9696	0.9397	0.8995	0.8704	0.8669	0.8540	0.8429
我们的模型	0.9850 ± 0.0001	0.9742 ± 0.0003	0.9487 ± 0.0023	0.9081 ± 0.0151	0.8812 ± 0.0089	0.8717 ± 0.0034	0.8717 ± 0.0021	0.8629 ± 0.0027

相关系数	电				汇率			
地平线	3.	6	12	24	3.	6	12	24
基于“增大化现实”技术	0.8845	0.8632	0.8591	0.8595	0.9734	0.9656	0.9526	0.9357
LRidge	0.8890	0.8594	0.8003	0.8806	0.9788	0.9722	0.9543	0.9305
LSVR	0.8888	0.8861	0.8961	0.8891	0.9782	0.9697	0.9546	0.9370
全科医生	0.8670	0.8334	0.8394	0.8818	0.8713	0.8193	0.8484	0.8278
伴随着	0.9402	0.9294	0.9202	0.9171	0.9759	0.9675	0.9518	0.9314
LSTNet-Skip	0.9283	0.9135	0.9077	0.9119	0.9735	0.9658	0.9511	0.9354
LSTNet-Attn	0.9243	0.9095	0.9030	0.9025	0.9717	0.9656	0.9499	0.9339
我们的模型	0.9429 ±0.0004	0.9337 ±0.0011	0.9250 ±0.0013	0.9133 ±0.0008	0.9790 ±0.0003	0.9709 ±0.0003	0.9564 ±0.0005	0.9381 ±0.0008

表 2 以 RAE、RSE 和 CORR 为指标的典型 MTS 数据集的结果。黑体字表现最佳;第二好的成绩用下划线表示。我们在 10 次运行中报告我们模型的平均值和标准差。除我们的模型结果外，所有数字均引用于 LSTNet 的论文[Lai 等人(2018)Lai, Chang, Yang, and Liu]。

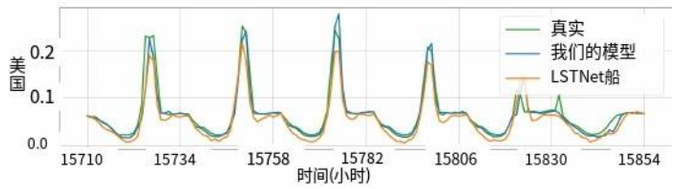

图 5 提出的模型与 LSTNet-Skip 在 3 小时视界流量测试集上的预测结果显然，所提出的模型在峰值后的平坦线周围和山谷中产生了更好的预测。

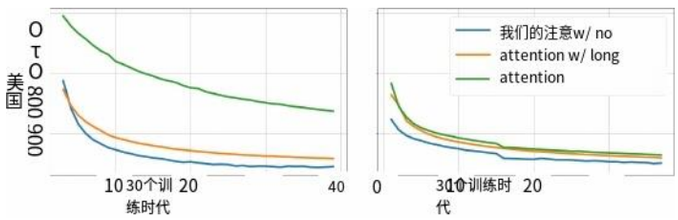

图 6 MuseData(左)和 LPD-5- cleaned(右)不同训练时段下的验证损失。

下一个指标是根相对平方误差(RSE):

$$\mathrm{RSE} = \frac{\sqrt{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n}(y_{t,i}-\hat{y}_{t,i})^2}}{\sqrt{\sum_{t=t_0}^{t_1}\sum_{i=1}^{n}(\hat{y}_{t,i}-\overline{\hat{y}_{t_0:t_1,1:n}})^2}}, \tag{18}$$

最后第三个指标是经验相关系数(CORR):

$$\mathrm{CORR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t=t_0}^{t_1}(y_{t,i}-\overline{y_{t_0:t_1,i}})(\hat{y}_{t,i}-\overline{\hat{y}_{t_0:t_1,i}})}{\sqrt{\sum_{t=t_0}^{t_1}(y_{t,i}-\overline{y_{t_0:t_1,i}})^2\sum_{t=t_0}^{t_1}(\hat{y}_{t,i}-\overline{\hat{y}_{t_0:t_1,i}})^2}}, \tag{19}$$

其中 y, y 在第 4.1 节中定义，y_t，$\forall\ t\in[t_0, t_1]$是测试数据的真值，y 表示集合 y 的均值。RAE 和 RSE 都不考虑数据尺度，分别是平均绝对误差(MAE)和均方根误差(RMSE)的归一化版本。RAE 和 RSE 越低越好，CORR 越高越好。

为了确定哪个模型在复调音乐数据集上更好，我们使用验证损失(负对数似然)、精度、召回率和 F1 分数作为测量值，这些测量值在复调音乐生成工作中被广泛使用[Nicolas Boulanger-Lewandowski 和 Vincent(2012)， Chuan 和 Herremans(2018)]。

	MuseData		
度规	精度	回忆	F1
W / o 的关注	0.84009	0.67657	0.74952
W/ Luong 注意	0.75197	0.52839	0.62066
W/建议注意事项	0.85581	0.68889	0.76333

	LPD-5-Cleansed		
度规	精度	回忆	F1
W / o 的关注	0.83794	0.73041	0.78049
W/ Luong 注意	0.83548	0.72380	0.77564
W/建议注意	0.83979	0.74517	0.78966

表 3 不同模型在复调音乐数据集上的准确率、召回率和 F1 分数。

6.5 典型 MTS 数据集的结果

在典型的 MTS 数据集上，我们使用 RAE/RSE/CORR 作为测试集的度量来选择验证集上的最佳模型。数值结果列在表 2 中，其中前两个表的度量是 RAE，其次是两个 RSE 度量表，最后是另外两个使用 CORR 度量的表。这两张表都表明，所提出的模型在所有数据集、视界和指标上的表现都优于几乎所有其他方法。此外，我们的模型能够处理各种数据集大小，从最小的 534 KB 汇率数据集到最大的 172 MB 太阳能数据集。结果表明，该模型在 MTS 预测方面具有一定的优越性。

与之前最先进的 LSTNet-Skip 和 LSTNet-Attn 方法相比，所提出的模型表现出优越的性能，特别是在包含最多时间序列的交通和电力方面。此外，在不存在重复模式的汇率方面，所建议的模型总体上仍然是最好的;LSTNet-Skip 和 LSTNet-Attn 的性能低于 AR、LRidge、LSVR、GP 和 SETAR 等传统方法。在图 5 中，我们还可视化并比较了所提出模型和 LSTNet-Skip 的预测。

总之，所提出的模型在周期性和非周期性 MTS 数据集上都达到了最先进的性能。

6.6 关于复调音乐数据集的结果

在本小节中，为了进一步验证所提出模型对离散数据的有效性和泛化能力，我们描述了在复调音乐数据集上进行的实验;实验结果如图 6 和表 3 所示。我们比较了三种 RNN 模型:LSTM, LSTM 与 Luong 注意，LSTM 与提出的注意机制。图 6 显示了跨训练时代的验证损失，在表 3 中，我们使用验证损失最低的模型来计算测试集上的精度、召回率和 F1 分数。

从结果中，我们首先验证了我们的说法，即典型的注意力机制在这样的任务上不起作用，就像在类似的超参数和

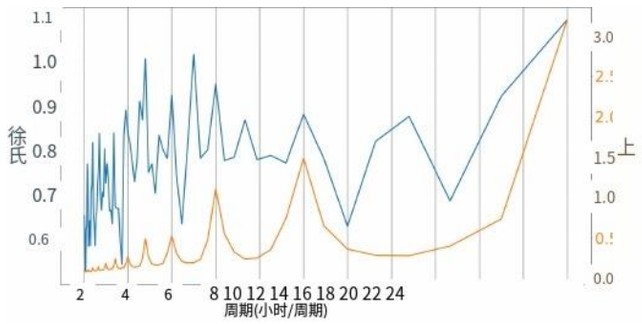

图 7(1)基于 3 小时视距的 Traffic 训练的 CNN 滤波器的 DFT 和(2)Traffic 数据集的每个窗口的幅度比较。为了使图更直观，横轴的单位是周期。

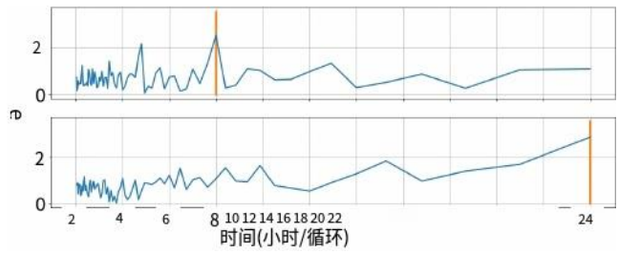

图 8 两种不同的 CNN 滤波器在 3 小时视域上训练，检测不同时期的时间模式。

可训练权值、LSTM 和所提出的模型优于这些注意机制。此外，与 LSTM 相比，该模型在整个学习过程中学习效率更高，在准确率、召回率和 F1 分数方面表现更好。

6.7 CNN 滤波器分析

DFT 是傅立叶变换(FT)的一种变体，它处理信号在时间上的等间隔采样。在时间序列分析领域，有大量的工作利用 FT 或 DFT 来揭示时间序列的重要特征 Huang and Liu(1998)， Bloomfield(1976)]。在我们的例子中，由于 MTS 数据也是等间隔和离散的，我们可以应用 DFT 来分析它。然而，在 MTS 数据中，有多个时间序列，因此我们自然地平均每个时间序列的频率分量的幅度，并得到一个单一的频域表示。我们把它表示为平均离散傅立叶变换(avg-DFT)。单频域表示揭示了 MTS 数据的主要频率成分。例如，可以合理地假设在

数据集	太阳能(Horizon = 24)			交通(地平线= 24)		
	位置	过滤器	W/o CNN	位置	过滤器	W/o CNN
Softmax	0.4397 ± 0.0089	0.4414 ± 0.0093	0.4502 ± 0.0099	0.4696 ± 0.0062	0.4832 ± 0.0109	0.4810 ± 0.0083
乙状结肠	0.4389 ± 0.0084	0.4598 ± 0.0011	0.4639 ± 0.0101	0.4765 ± 0.0068	0.4785 ± 0.0069	0.4803 ± 0.0104
Concat	0.4431± 0.0100	0.4454± 0.0093	0.4851± 0.0049	0.4812±0. 0082	0.4783±0. 0077	0.4779±0. 0073

数据集	电力(Horizon = 24)			MuseData		
	位置	过滤器	W/o CNN	位置	过滤器	W/o CNN
Softmax	0.0997 ± 0.0012	0.1007 ± 0.0013	0.1010 ± 0.0011	0.04923 ± 0.0037	0.04929 ± 0.0031	0.04951 ± 0.0041
乙状结肠	0.1006 ±0.0015	0.1022 ±0.0009	0.1013 ±0.0011	0.04882 ±0.0031	0.04958 ±0.0028	0.04979 ±0.0027
Concat	0.1021± 0.0017	0.1065± 0.0029	0.1012± 0.0008	0.05163± 0.0040	0.05179± 0.0036	0.05112± 0.0027

表 4 消融研究。太阳能、交通和电力的评价指标为 RSE, MuseData 的评价指标为负对数似然。我们报告十次运行的平均值和标准差。在每个语料库上，粗体文本代表最好的，下划线文本代表第二好的。

图 5，由图 7 所示 Traffic 数据集的 avg-DFT 验证。

由于我们期望我们的 CNN 滤波器能够学习时间 MTS 模式，因此平均 CNN 滤波器中的主要频率成分应该与训练 MTS 数据的频率成分相似。因此，我们还对 k = 32 个 CNN 滤波器应用 avg-DFT，这些滤波器是在 3 小时范围内的流量上训练的;在图 7 中，我们将结果与 Traffic 数据集的每个窗口的 avg-DFT 一起绘制。令人印象深刻的是，两条曲线在大多数情况下都在同一时间段达到峰值，这意味着学习到的 CNN 滤波器类似于 DFT 中的基。在 24 小时、12 小时、8 小时和 6 小时的时间段内，不仅流量数据集的大小达到峰值，CNN 过滤器的大小也达到峰值。此外，在图 8 中，我们展示了不同的 CNN 过滤器的行为不同。一些专门捕捉长期(24 小时)的时间模式，而另一些擅长识别短期(8 小时)的时间模式。总的来说，我们建议所提出的 CNN 滤波器在 DFT 中起到基的作用。正如[Rippel 等人(2015)Rippel, Snoek, and Adams]的工作所证明的那样，这样的“频域”可以作为 CNN 在训练和建模中使用的强大表示。因此，LSTM 依赖于所提出的注意机制提取的频域信息来准确预测未来。

6.8 消融研究

为了验证上述改进来自每个添加的组件，而不是特定的一组超参数，我们对太阳能、交通、电力和 MuseData 数据集进行了消融研究。有两种主要设置:一种控制我们如何关注 RNN 的隐藏状态 H，另一种控制我们如何将评分函数 f 集成到提议的模型中，甚至禁用该函数。首先，在提议的方法中，

我们让模型关注每个位置(H_i^C)上各种滤波器的值;我们也可以考虑关注相同滤波器在不同位置($(H^C)_{>i}$)或 H 的行向量($H_{i>}$)上的值。这三种不同的方法对应于表 4 中的列标题:“Position”、“Filter”和“Without CNN”。其次，在典型的注意力机制中，通常对评分函数 f 的输出值使用 softmax 来提取最相关的信息，而我们使用 sigmoid 作为我们的激活函数。因此，我们对这两种不同的函数进行比较。另一种可能的预测结构是将之前的所有隐藏状态连接起来，让模型自动学习哪些值是重要的。考虑到这两组设置，我们在这四个数据集上训练了具有所有可能结构组合的模型。

MuseData 结果表明，具有 s 形激活和 H_i^C (位置)关注的模型明显是最好的，这表明该模型具有合理的预测效果。无论从模型中去除哪个提议的成分，性能都会下降。例如，使用 softmax 而不是 sigmoid 将负对数似然从 0.04882 提高到 0.04923;如果我们不使用 CNN 滤波器，我们会得到一个更糟糕的模型，其对数似然值为负 0.4979。此外，我们注意到，在表 4 中的前三个数据集(太阳能、交通和电力)上，所提出的模型与使用 softmax 的模型之间没有显著的改进。考虑到我们使用 sigmoid 的动机，如第 4.3 节所述，这并不奇怪。最初，我们期望 CNN 过滤器找到基本模式，并期望 sigmoid 函数帮助模型将这些模式组合成一个有帮助的模式。然而，由于这三个数据集具有很强的周期性，有可能使用少量的基本模式就足以进行良好的预测。然而，总的来说，所提出的模型更加通用，并且在不同的数据集上产生稳定和有竞争力的结果。

7 结论

在本文中，我们专注于 MTS 预测，并提出了一种新的时间模式注意机制，该机制消除了典型注意机制在此类任务上的局限性。我们允许注意维度具有特征智能，以便模型不仅在同一时间步长内，而且在所有以前的时间和序列中学习多个变量之间的相互依赖关系。我们在玩具示例和现实世界数据集上的实验强烈支持这一想法，并表明所提出的模型达到了最先进的结果。此外，过滤器的可视化也以一种更容易被人类理解的方式验证了我们的动机。

参考文献

A. Krizhevsky 和 Hinton(2012)。A. Krizhevsky IS, Hinton GE(2012)基于深度卷积神经网络的 Imagenet 分类。神经信息处理的进展-

ing 系统 pp 1097-1105

Bahdanau 等人(2015)Bahdanau、Cho 和 Bengio。Bahdanau D, Cho K, Bengio Y(2015)联合学习对齐和翻译的神经机器翻译。ICLR

布卢姆菲尔德(1976)。BloomfieldP(1976)《时间序列的傅立叶分析:导论》。约翰威利

布查奇亚与布查奇亚(2008)。布查奇亚 A, 布查奇亚 S(2008)时间序列预测的集成学习。第一届非线性动力学与同步国际研讨会论文集

曹和 Tay(2003)。曹立军，Tay FEH(2003)基于自适应参数的支持向量机在金融时间序列预测中的应用。IEEE 汇刊第 1506-1518 页

陈等人(2008)陈、王和 Harris。陈生，王 XX, Harris CJ(2008)基于正交最小二乘基寻优的非线性系统辨识。IEEE 汇刊控制系统第 78-84 页

川和 Herremans(2018)。Chuan CH, Herremans D(2018)基于深度网络的复调音乐时间调性关系建模。URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16679

达斯古普塔和奥索伽米(2017)。Dasgupta S, 奥索伽米 T(2017)时间序列预测的非线性动态 Boltzmann 机

大卫 E.鲁梅尔哈特和 Williams(1986)。David E 鲁梅尔哈特 GEH, Williams RJ(1986)反向传播误差学习表征。Nature 第 533-536 页

Elman(1990)。Elman JL(1990)在时间中寻找结构。认知科学第 179-211 页

Frigola and Rasmussen(2014)。Frigola R, Rasmussen CE(2014)高斯过程贝叶斯非线性系统辨识的集成预处理。IEEE Conference on Decision and Control 第 552-560 页

Frigola-Alcade(2015)。Frigola-Alcade R(2015)高斯过程的贝叶斯时间序列学习。博士论文，英国剑桥大学

G. E. Box 和 Ljung(2015)。G E Box GCR GM Jenkins, Ljung GM(2015)时间序列分析:预测与控制。John Wiley & Sons

张、胡(1998)。张，胡明(1998)基于人工神经网络的预测研究进展。《国际预测杂志》第 35-62 页

董浩文，杨(2018)。董浩文，肖文毅，杨勇(2018)MuseGAN:符号音乐生成与伴奏的多音轨序列生成对抗网络

Hochreiter and Schmidhuber(1997)。李晓明，李晓明，李晓明，等(1997)长短时记忆。神经计算 9(8):1735-1780,DOI 10.1162/neco.1997.9.8。1735, URL https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10。1162/neco.1997.9.8.1735

J.康纳和马丁(1991)。J 康纳 LEA, 马丁 DR(1991)循环网络和 NARMA 建模。《神经信息处理系统进展》pp 301-308

贾恩和 Kumar(2007)。Jain A, Kumar AM(2007)水文时间序列预测的混合神经网络模型。应用软计算(Applied Soft Computing) 7(2): 585-592

J.Werbos(1990)。JWerbos P(1990)穿越时间的反向传播:它做什么以及如何做。IEEE Proceedings of the IEEE pp 1550-1560

金(2003)。Kim KJ(2003)基于支持向量机的金融时间序列预测。Neurocomputing 55 (1): 307 - 319

Kyunghyun Cho 和 Bengio(2014)。Kyunghyun Cho DB Bart Van Merrienboer, Bengio Y(2014)关于神经机器翻译的特性:编码器-解码器方法。arXiv 预印 arXiv:14091259

赖等(2018)赖，常，杨，刘。赖刚，常万维，杨勇，刘辉(2018)基于深度神经网络的长、短期时间模式建模。SIGIR, pp 95-104

LeCun 和 Bengio(1995)。LeCun Y, Bengio Y(1995)图像、语音和时间序列的卷积网络。脑理论与神经网络手册

Luong 等人(2015)Luong, Pham 和 Manning。Luong T, Pham H, Manning CD(2015)基于注意的神经机器翻译的有效方法。2015 年自然语言处理经验方法会议论文集，pp 1412-1421

黄、刘(1998)。黄 NE，刘辉(1998)非线性非平稳时间序列分析的经验模态分解和 Hilbert 谱。Proc Roy Soc, 2012,44 (4):903 - 995

Nicolas Boulanger-Lewandowski 与 Vincent(2012)。Nicolas Boulanger-Lewandowski YB, Vincent P(2012)高维序列的时间依赖性建模:在复调音乐生成和转录中的应用

秦等(2017)秦、宋、程、程、江、Cottrell。秦勇，宋迪，程辉，程伟，江光，Cottrell GW(2017)基于双阶段注意力的递归神经网络时间序列预测。In: IJCAI’17,pp 2627-2633, URL http://dl.acm。org/citation.cfm ? id = 3172077.3172254

Raffel(2016)。Raffel C(2016)基于学习的序列比较方法，应用于 audio-to-MIDI 对齐和匹配。博士论文

Rippel 等人(2015)Rippel, Snoek 和 Adams。Rippel O, Snoek J, Adams RP(2015)卷积神经网络的谱表示。NIPSpp 2449-2457

S.罗伯茨和 Aigrain(2011)。S 罗伯茨 MESRNG M Osborne, Aigrain S(2011)高斯过程的时间序列建模。Phil Trans R Soc A

Sutskever 等人(2014)Sutskever、Vinyals 和 Le。Sutskever I, Vinyals O, Le QV(2014)基于神经网络的序列到序列学习。《神经信息处理系统进展》pp 3104-3112

Tong 和 Lim(2009)。Tong H, Lim KS(2009)阈值自回归，极限环和周期数据。见:《对非线性世界的探索:Tong 对统计学的贡献赏析》，《世界科学》，第 9-56 页

V.万普尼克(1997)。V 万普尼克 ASea S E Golowich(1997)支持向量法的函数逼近，回归估计，和信号处理。《神经信息处理系统进展》pp 281-287

张(2003)。张 gp(2003)使用 ARIMA 和神经网络混合模型的时间序列预测。《神经计算》(Neurocomputing)第 159-175 页