2023年研究生机器学习课程项目

提交日期:二零二四年一月十日中午12点前（实际完成需至少一个月,视调试经验或有長短）。以邮件提交时间为准,逾期恕不接收。

合作方式:可以至多三人组队,请在报告上写明各位作者的贡献。且请注明各自所在的研究领域。队内讨论为主,但不阻碍队间交流,但从其他队凡有借鉴之处,须在参考文献中注明,否则视为剽窃。

问题的来源:现代的电网需要根据时序变化的需求来把电力分配到不同用户区域。但实际中,要预测特定用户 区域的未来需求很困难,因为它随工作日/假日、季节、天气、温度等的不同因素变化而变化。由于当前没有一种有效的方法来预测未来的用电量,电力公司就不得不根据经验值做出决策,而经验值的阈值通常远高于实际需求。保守的策略导致不必要的电力和设备折旧浪费。值得注意的是,变压器的油温可以有效反映电力变压器的工况。因此预测变压器的油温可以用来研究电力变压器极限负载能力,设法避免不必要的浪费。

我们现在就要通过一个电力变压器数据集ETTh1来预测电力变压器的油温。这个数据集提供了两年的记录,每个数据点每小时记录一次。数据集包含2年X 365天X 24小时= 17520个数据点。每个数据点均包含8维特征， 包括数据点的记录日期、预测值 “油温”以及6个不同类型的外部负载值。所有的数据都经过了预处理,并且以.csv的格式存储。

具体数据形式如下：

	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
0	2016-07-01 00:00:00	5.827	2.009	1.599	0.462	4.203	1.340	30.531000
1	2016-07-01 00:15:00	5.760	2.076	1.492	0.426	4.264	1.401	30.459999
2	2016-07-01 00:30:00	5.760	1.942	1.492	0.391	4.234	1.310	30.038000
3	2016-07-01 00:45:00	5.760	1.942	1.492	0.426	4.234	1.310	27.013000
4	2016-07-01 01:00:00	5.693	2.076	1.492	0.426	4.142	1.371	27.787001

数据中各列的含义：

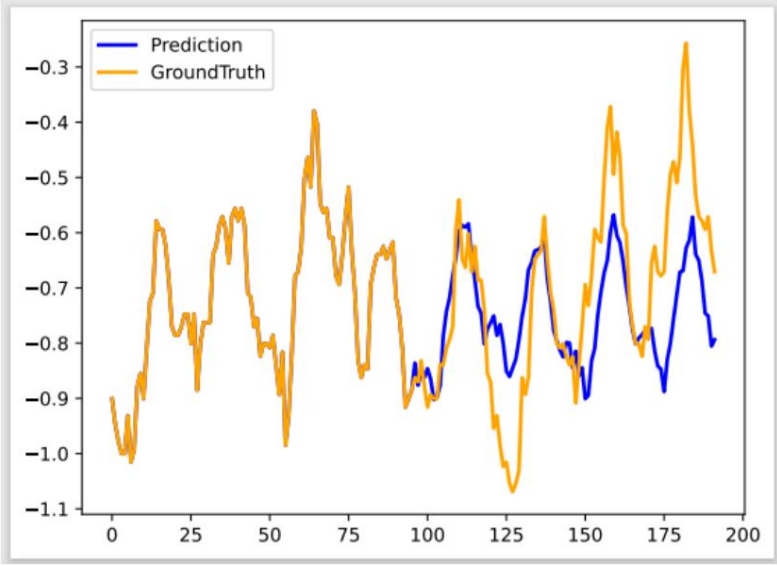
Field	date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
Description	The recorded **date**	High UseFul Load	High UseLess Load	Middle UseFul Load	Middle UseLess Load	Low UseFul Load	Low UseLess Load	Oil Temperature (target)

此数据集在此下载并可进一步获得数据的详情： https://github.com/zhouhaoyi/ETDataset

预测任务:为根据过去I=96小时的曲线来预测将来（i) O=96小时（見示意图)和（ii) O=336小时两种長度的变化曲线（须分别训练,即長程预测的模型参数不能用来进行短程预测）。注意这是多变量预测问题,输入 是:样本数X 输入序列長度X 变量数,输出是:样本数X 预测序列長度X 变量数。按方法分为三部分。

前两部分为基础题,第三部分为开放题,各占三分之一： 1. 用
LSTM模型预测， 2. 用
Transformer模型预测， 3. 用
自己提出的改进模型预测,结构不限,此部分以原理的新颖程度为第一、性能为第二的标准评分。

训练与测试:Train/Val/Test划分为6:2:2。数据集采用滑动窗口制作,请参考教程。请用两种标准测试,即
MSE与MAE。 至少五轮结果平均,并提供std。



预测结果示意图（输入96小时,预测長度96小时）。仅显示了最后一个变量即油温的曲线,实际需预测七个变量。

提交方式:实验报告由1.问题介绍、2.模型（可以包含少量伪代码）、3.结果与分析、4.讨论四部分构成,同时提交代码（可以给出Github)。结果须截图贴在报告里,并画出油温预测与Ground Truth曲线的对比。请注意三种方法之间的比较。如果自行提出的方法虽新颖而性能欠佳,但原因分析有力,同样可以得到较佳的分数。务请注明参考文献,否则视为剽窃,每剽窃一处扣33分。允许用ChatGPT一类工具写报告,但仅限于撰写并请注明,必要处参考文献仍不可缺。

参考和提示:对于前两个部分,我们提供了几个Tutorials。其中以单变量为主,我们的数据是多变量的。对第三部分我们提供了一些近年有代表性的参考文章,并作了简介。但请注意不允许照抄其中的模型,必须自行设计。时间序列预测和其它机器学习问题,如语言处理、语音辨识、基因序列预测等有很多相似之处,可以充分借鉴这些领域的方法,但同时要注意时间序列的特点,尤其是数据本身的特征。值得提醒的是,许多非
Transformer型模型有卓越的表现,且训练时间远短于Transformers,如卷积类的模型。

以上是一个有实际应用场景的问题,可以为电力工业提供切实的工具。希望各位在这个课程项目中有所获益， 或可预测自家用电情况，或可在电力公司谋一份美差。 Have fun!!

参考教程（第1、2题）：

[1] https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/ [2] https://weibaohang.blog.csdn.net/article/details/128595011
[3]https://towardsdatascience.com/how-to-use-transformer-networks-to-build-a-forecasting-model-297f9270e630

参考文章（第3题）：

[1] Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]// Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 11106- 11115. 基于Transformer改进的长序列预测模型,解决了二次时间复杂度、高内存占用以及编码器-解码器架构的固有限制的问题。主要提出ProbSparse自注意力机制,利用稀疏注意力在时间复杂度和内存使用方面实现了O(L log L)。

[2] 吴华,徐静,王静,等。 Autoformer:具有长期自相关性的分解变压器

系列预测[J].神经信息处理系统的进展,2021,34:22419-22430。
基于Transformer的具有自相关机制的分解架构。主要提出了基于序列周期性的自相关机制,在子序列级别进
行依赖关系发现和表示聚合,以及将序列分解嵌入到深度模型中。

[3]周涛,马Z,文Q,等。 Fedformer:用于长期序列预测的频率增强分解变压器[C]//机器学习国际会议。 PMLR,2022 年:
27268-27286。
基于Transformer的频率增强分解模型,主要利用大多数时间序列在傅里叶变换中往往具有稀疏表示的事实,
提出了频率增强模块,以及与季节趋势分解方法相结合。

[4]曾安,陈明,张立,等。 Transformer 对于时间序列预测有效吗?[C]//AAAI 人工智能会议论文集。 2023, 37(9): 11121-11128。

质疑Transformer的有效性,认为排列不变自注意力机制的本质不可避免地会导致时间信息丢失。主要引入了
一组极其简单的单层线性模型来实现并比较。

[5]张Y,严J. Crossformer:利用跨维度依赖进行多元时间序列预测的Transformer[C]//第十一届学习表示国际会议。 2022 年。

在现有的基于Transformer 的模型主要侧重于对时间依赖性（跨时间依赖性)进行建模的基础上,再关注不同
变量之间的依赖性（跨维度依赖性）,从而保留时间和维度信息。

[6] 聂Y,Nguyen NH,Sinthong P,等。一个时间序列相当于 64 个字:使用 Transformer 进行长期预测[J]. arXiv 预印本
arXiv:2211.14730, 2022。
主要提出将时间序列分割为子序列级别的补丁,作为Transformer 的输入标记,以及多变量时间序列的通道独
立性,其中每个通道包含单个单变量时间序列。

[7]吴华,胡涛,刘Y,等。 Timesnet:用于一般时间序列分析的时间二维变化建模[J]。 arXiv 预印本 arXiv:2210.02186, 2022。

基于时间序列的多周期性,将复杂的时间变化分解为多个周期内和周期间的变化。主要提出将一维时间序列转
换为一组基于多个周期的二维张量,将时间变化的分析扩展到二维空间。

（谢谢N.Bian同学的大力协助。）