测试:文本原型对齐嵌入激活LLM时间序列能力

Chenxi Sun1,2, Yaliang Li3 , Hongyan Li1,2, Shenda Hong4,5 1北京大学智能科学与技术学院2机器感知教育部重点实验室,北京大学3阿里巴巴集团4国家健康数据科学研究所、北京大学医学技术研究所、北京大学医学部 sun

chenxi@pku.edu.cn, yaliang.li@alibaba-inc.com, leehy@pku.edu.cn , hongshenda@pku.edu.cn

抽象的

这项工作总结了使用当今语言模型（LLM)完成时间序列（TS)任务的两种策略:LLM- for-TS,为 TS 数据设计和训练一个基本的大模型； TS-for-LLM,使预训练的LLM能够处理TS数据。

考虑到数据积累不足、资源有限和语义上下文要求,本工作重点研究TS-for-LLM方法,旨在通过设计适合LLM的TS嵌入方法来激活LLM对TS数据的能力。所提出的方法被命名为TEST。它首先对 TS 进行标记,构建一个编码器通过实例明智、特征明智和文本原型对齐对比来嵌入它们,然后创建提示以使 LLM 对嵌入更加开放,最后实现 TS 任务。使用8个不同结构和规模的LLM对TS分类和预测任务进行了实验。虽然其结果不能显着优于当前为TS任务定制的SOTA模型,但通过将LLM视为模式机,它可以赋予LLM处理TS数据的能力,而不会影响语言能力。本文旨在作为基础工作,激发进一步的研究。

介绍

实施基于时间序列 (TS) 的任务（例如医疗、工业和气象)是一个研究密集型领域。
相关模型从统计模型发展到RNN、 CNN、Transformers。如今,我们看到大规模预训练语言模型（LLM)在 NLP 和 CV 领域的快速增长和卓越表现。

因此,询问是否有可能将 TS 与 LLM 整合起来似乎是有意义的。然而,根据实验,大多数法学硕士在抽象时间序列方面并没有取得重大进展。在这项工作中,我们设想了两种实现TS+LLM 范式的方法：

- TS 法学硕士。针对 TS 数据,从头开始设计和预训练一个基本的大模型,然后针对各种下游任务相应地微调模型。
- 法学硕士 TS。基于现有的LLM，使其能够处理TS数据和任务。与其创建一个新的LLM,不如设计一些机制来为LLM定制TS。

我们承认第一种方式是最根本的解决办法,因为预训练是灌输的关键步骤模型的知识。而第二种方式实际上很难突破模型原有的能力。

然而,在这项工作中,出于以下三个考虑,我们仍然关注第二种方式：

数据。 LLM-for-TS需要大量的积累数据。
由于TS更专业,且涉及隐私,因此相比文本或图像数据更难大量获取,尤其是在非工业岗位； TS-for- LLM 可以使用相对较小的数据集,因为其目标仅仅是协助现有的 LLM 推断 TS。

模型。 LLM-for-TS 专注于垂直行业。由于跨领域的 TS 存在巨大差异,因此必须从一开始就建立和训练针对医疗 TS、工业 TS 等的各种大型模型； TS-for-LLM 需要很少甚至不需要培训。通过插件模块的使用,使使用更加通用、方便。

用法。 LLM-for-TS 适用于涉及专家的情况； TS-for-LLM 保留了 LLM 的文本功能,同时提供丰富的补充语义、易于访问且用户友好。

基于预训练的LLM,最自然的方法是将TS视为文本数据。例如,可能的对话是： [Q] 通过以下平均动脉压序列（以毫米汞柱为单位)诊断患者是否患有脓毒症:88、95、78、65、52、30。 [A] 是。然而,TS 通常是多变量的,而文本是单变量的。例如,诊断脓毒症时,除了平均动脉压外,还需要包括心率、乳酸等数十种生命体征和实验室值。处理单变量文本的LLM会将一个多变量TS转化为多个单变量序列并一一输入。然而,这会导致三个缺点。

首先,不同的提示、顺序、连接语句会产生不同的结果;其次,较长的输入序列可能会使LLM效率低下并且难以记住之前的单变量TS;第三， TS 中多元依赖性的关键方面将被忽略。

因此,在这项工作中,我们对 TS 进行标记,设计一个模型来嵌入 TS 标记,并替换 LLM 的嵌入层。这种方式的核心是创建LLM可以理解的嵌入。

TS 嵌入可以通过包含典型属性、关联属性和依赖属性来提供身份,并揭示相关系统的底层机制。高质量

嵌入对于下游任务至关重要,它们可以用作深度学习模型可以理解的计算表型。例如,脓毒症诊断模型可以使用 ICU 患者的生命体征嵌入。

使嵌入可以被语言模型理解。大多数多模式方法都使用对齐方式。例如， SOTA 方法通过图像的文本描述来对齐文本嵌入和图像嵌入。然而,与其他数据模态相比,TS 缺乏视觉线索,并且由于其复杂的特性而存在注释瓶颈。

只有少数特定的TS,例如ECG,每段都有文字描述,可以实现图文匹配路线。但在大多数情况下,这是不可行的。

自监督对比学习可以通过利用内在信息而不是依赖预先定义的先验知识来设计借口任务,从而避免注释瓶颈。对比学习采用实例辨别借口任务来使相似对更接近,同时在嵌入空间中将不相似对推开。目前,已经努力实现实例级对比度、时间级对比度和原型级对比度,并取得了可喜的结果。这些方法通过后续分类、预测或聚类模型（例如 SVM)来评估 TS 嵌入的有效性。然而,这些简单且新训练的模型与复杂且预先训练的法学硕士有很大不同。无约束对比学习生成的表示向量可能会大大偏离法学硕士的认知嵌入空间。

为了解决这个问题,我们提出了一种 TimE 系列代币的嵌入方法,以对齐LLM (TEST)的文本嵌入空间。在比较学习的基础上， TEST使用正交文本嵌入向量作为原型来约束TS的嵌入空间,并通过识别特征原型来突出模式以激活LLM的模式机能力。这项工作的贡献是：

·总结两种TS+LLM 范例:LLM-for-TS 和TS-for-LLM,以及潜在的实施方式。 ·提出TEST以实现LLM 的TS。

TEST可以为 TS 令牌生成基于相似性、实例明智、特征明智和文本原型对齐的嵌入。 · TS分类和预测任务的实验表明， TEST可以激活LLM的TS能力。它可以将原始LLM产生的随机和不令人满意的结果提升到基线。

正如TEST的名称所暗示的,虽然 TS-for-LLM 不能明显优于当前为 TS 任务定制的 SOTA 模型,但它是一个前瞻性的测试,我们希望能为未来的研究奠定基础。事实上,它确实赋予了法学硕士新的能力,并体现了其作为模式机器的品质。

相关工作

TS+法学硕士。现有的做法有以下三种。

PromptCast (Xue and Salim 2023) 和 Health Learner (Liu et al. 2023) 将 TS 视为文本序列。他们直接将单变量数值TS输入LLM并设计提示执行TS任务。显然,它们受到引言部分总结的三个困境的限制。 METS (Li et al. 2023) 对齐临床报告文本和心电图信号。它满足 ECG-文本多模态条件,但不能推广到大多数没有分段注释的 TS 数据。

TS 嵌入。我们利用对比学习的思想。

它是以实例歧视为借口任务的一种特殊形式的自监督学习。大多数方法侧重于实例级对比,独立处理实例（全长TS、TS 片段等）,并将锚点的增强视图视为正视图,将其余视图视为负视图（Chen 等人,2020）。

在TS对比学习中,考虑到固有的时间依赖性,研究人员探索了在细粒度时间水平上区分上下文信息的可行性（Meng et al. 2023）。实例的选择包括系列间和系列内（Woo et al. 2022;Zheng et al. 2023）。这种时间级别对比的想法（Franceschi、 Dieuleveut 和 Jaggi 2019)为我们提供了嵌入 TS 令牌的基本建模方法。

然而,直接对比无法弥合 TS 嵌入和 LLM 的可理解空间。在我们的设置中,我们更愿意冻结预先训练的 LLM,并让嵌入妥协。也就是说,我们使用LLM中的文本标记嵌入来限制和引导TS标记嵌入。

我们注意到LLM的本质实际上是通用模式机（Mirchandani et al. 2023）。因此,无论token列表的组合是否具有人类可以理解的语义,我们都会强制对齐TS token的模式和文本token的模式。也就是说,TS令牌列表可以近似地由没有语义信息的句子来表达。但我们的目标是获得LLM可以理解的模式序列。

受到原型级对比（Caron et al. 2020)的启发,它超越了独立性假设并利用了样本中存在的潜在聚类信息。我们可以选择一些文本嵌入作为基本原型来引导学习。然而,除了对齐之外,我们还需要考虑原型选择、差异化（Meng et al. 2023）、均匀性（Wang and Isola 2020）、稳定性（Huang et al. 2023)等问题。

方法

概述

为了实现 LLM的TS ,我们的方法有两个关键步骤： ·图 1:对 TS 进行标记并训练编码器来嵌入 TS 具有对比学习的标记。

·图2:创建提示以使LLM 对嵌入更加开放并实施TS 任务。

TS Token 增强和编码定义 1（时间序列的 Token Embedding)多变量时间序列x = {x有D个变量和T个时间点。可以通过分割函数Fs : $x_{t:t+s,d}$ 将其分割为K 个不重叠子序列s = {sk}的列表,其中sk = xti:tj的长度是任意的， 1 ≤ ti < tj ≤ T。我们称s作为 token 列表

$x_{t:t+s,d}$ 时间,维度 $t=1,d=1$ $k=1$

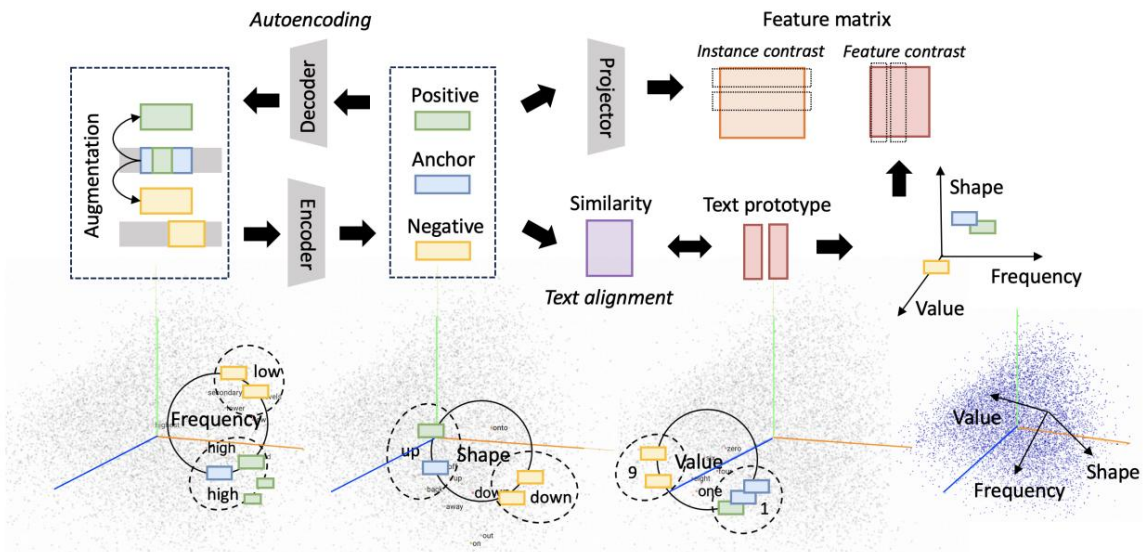

图 1:通过实例感知和特征感知对比学习进行文本原型对齐的时间序列嵌入

时间序列x.此外,每个标记可以通过嵌入函数→ ek ∈ RM嵌入到M维表示空间。最后,令牌嵌入

Fe : sk ∈ R x 的列表深×厚

是 e = {ek} Kk=1 = Fe(s) = Fe(Fs(x))。

首先,我们将 TS s = Fs(x) 标记化。在 TS 表示学习中,分割函数Fs通常是滑动窗口（Yue et al. 2022）。我们使用随机长度和步长来分割 TS,并得到许多 token。

然后,我们定义一个 TS token s作为锚实例。

对于正数s +,我们定义两个采集源。第一个是重叠实例,使用与 s 具有相同子序列的实例。第二个是增强实例,从增强系列 Tweak、 sstrong Tstrong 中产生两个视图。具体来说,为了获得弱增强视图xweak,我们使用抖动和缩放策略,向信号添加随机变化并放大其幅度。为了获得强弱的增强视图sstrong,我们使用排列和抖动策略,将序列分成随机数量的片段并随机打乱它们。

对于负数s −,我们使用与 s 不具有相同子序列的实例。

获得锚正负后,我们构建了一个神经网络作为编码器,将实例嵌入到向量e = Fe(s) 中。编码器Fe必须能够从 TS 中提取相关信息,需要在训练和测试方面都具有时间和内存效率,并且必须允许可变长度输入。因此,受到（Franceschi、Dieuleveut 和 Jaggi 2019)的启发,我们使用指数扩张的因果卷积网络来处理 TS。我们将网络的每一层构建为因果卷积、权重归一化、激活和残差连接的组合。每层都被赋予一个指数增加的膨胀参数。

我们还使用自动编码训练了解码器fd

验证的有效 1/氮 尼 =1 sim(s, fd(e))确保表征损失Lae =嵌入和后续性。因为我们的主要目标是检索编码器,该解码器

同样可以在不损害未来进程的情况下被拆除。

实例方面和特征方面的对比

基本的实例对比学习独立对待每个实例,并设计实例辨别借口任务,以保持相似实例靠近,不同实例远离。

我们将同一实例的增强视图视为唯一的正对,并将B大小的小批量中的所有剩余视图视为负对。具体来说,我们应用了对比学习的 MoCo (He et al. 2020) 变体,它利用动量编码器来获取正对的表示,并使用带有队列的动态字典来获取负对。给定一个 TS token 实例和B 个负样本,实例方面的对比损失如公式 1 所示。在给定一个实例嵌入e 的情况下,我们为对比损失选择一个随机时间步t并构造一个投影头fp,这是一层MLP,得到+/-是对应的in-fp(e)的正/负。 e动量编码器的姿态。 σ(e, e+/-)用于通过类似余弦相似度的相似度函数sim来计算两个投影向量fp(e)和fp(e +/-)之间的相似度。 τI是实例级温度参数。

$$\mathrm{Lins} = -\log \frac{\exp(\sigma \mathrm{ins}(e, e+))}{\exp(\sigma \mathrm{ins}(e, e+)) + \exp(\sigma \mathrm{ins}(e, e-))} \quad \sigma \mathrm{ins}(e, e+/-) = \frac{\mathrm{sim}(fp(e), fp(e+/-))}{\tau I} \tag{1}$$

然而,实例对比学习倾向于将语义相似的样本视为否定样本（Caron et al. 2020）。

为了解决这个限制,我们提出了特征对比来打破实例之间的独立性。

如图1所示,embedding后,由实例的表示向量构成特征矩阵

右宽×米

在一个小批量中。其中每一行都是一个实例的嵌入,因此行可以被视为方程1中使用的实例的软标签。除了行之外,特征矩阵的列也具有语义信息。 （Li et al. 2021)提出,当投影期间表示的维数与簇的数量相匹配时,由于观察到标签作为表示,所以列可以进一步被视为簇表示。但这种聚类方法需要先验知识来预先指定聚类的数量,这对于本工作中未标记的 TS 数据来说并非易事。

因此,我们建议将列视为特征的软标签,并对相似特征组之间进行区分。这里我们得到了anchor实例的两种正特征矩阵和负特征矩阵mweak+， mstrong+， m－ ,其中m+/－ = Fe(e+/－)， m+/－ ∈ R B×M， e +/- = {ei}

B
i = 1。 我们缩小了两个正特征矩阵mweak+、 mstrong+的相同列之间的差距,如等式 2 第一项所示,并放大了正特征矩阵m+和负特征矩阵m－之间的差距,如第二项所示。我们将矩阵中的列标记为 m ∈ mT

。

中号
叶酸 = - σfea (mweak+ , 强+米我)
mi∈mT,i=1
特征对齐

中号
+ σfea (m+我，) m－
mi∈mT,i=1
特征差异

中号 ,MW+))
⇒ － 我=1 日志 exp(sim(毫秒+我 τF
,米+)
M j=1[exp(sim(m+ i τF) + exp(sim(m+ i τF ,m－)j)]
特征类别一致性

(2)

如公式2所示,特征对比主要是在正负之间对齐和区分相同的特征列。然而,由于相似性占主导地位,这可能会导致表示空间在一个小区域内收缩。我们发现确保功能之间的差异可以更好地解决这个问题。也就是说,我们在不同特征列之间添加对比学习,如修改后的等式2所示。它确保正数mw +的每个特征列相似,同时扩大不同特征列m+ m+ m－ j , j之间的差异，

我 , 女士+ 我 我，

其中τF是特征级温度参数,用于增加差异并控制负样本判别力。请注意,我们的特征对比度损失的形式与 infoNCE 的形式类似（He et al. 2020）。

更重要的是,特征列差异的注入也可以极大地帮助后续实现文本原型对齐对比。因为这种对比会将选定的文本标记应用于特征列,如坐标轴。

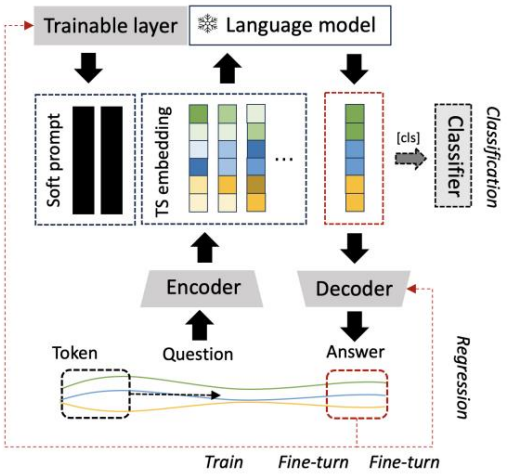

图2:TS任务的LLM框架

文本原型对齐对比为了使 LLM 理解 TS 嵌入,我们将其与文本表示空间对齐。

预训练的 LLM 有自己的文本标记嵌入。例如,小型、中型、大型 GPT-2 将单词词典中的文本标记分别嵌入到768、1024 和 1280 维的表示空间中（Radford 等人，2017）。2019）。天真的,如等式3的第一项所示,我们可以使用相似度估计强制对齐TS token e和文本token tp的这两种嵌入向量。

例如,虽然 TS token 缺乏文本注释,但我们可以将它们的嵌入放置在 TS 的典型文本描述附近,例如值、形状和频率,如图 1 所示。通过这种方式,直观地期望各种 TS token可以代表各种描述性术语,如小、大、上、下、稳定、波动等。当然,上面的例子是基于最近邻原则的,因为文本标记的嵌入空间是离散的,类似于向量表,但我们的 TS 标记的嵌入空间是连续的。

Ltext = － sim(tp, e) + Lf ea(e · tp, e+ · tp, e－ · tp)
文本对齐 文字对比

(3)

然而,当然,实际结果不会符合我们的预期,因为我们没有提供受监督的标签或基本事实。例如,具有上升趋势的子序列的嵌入可能非常接近于下降文本的嵌入,甚至是不描述趋势的文本的嵌入。但语义能否被我们理解并不重要。与往常一样,事实是人类无法理解模型的感知模式。最近,研究人员证明了法学硕士是模式机（Mirchan-dani et al. 2023）。因此,在这项工作中,我们实现了“TS→令牌列表→嵌入列表→文本模式列表”来激活LLM执行TS任务的能力。例如,通过简单的提示， LLM可以识别不同类模式之间的差异,从而实现TS分类任务。

这样,文本原型的选择可以放宽,不一定是与TS数据相关的描述。在这项工作中，

为了使两个嵌入空间更好地匹配,受到多语言单词嵌入的无监督对齐（Alaux 等人,2019)的启发,我们将 TS 嵌入映射到枢轴文本嵌入。具体来说,首先,我们通过向量的相似度约束,保证两个空间的范围大致相同,最大化TS向量与文本原型向量之间的余弦相似度,如式3第一项所示;其次,我们使用文本原型作为坐标轴来映射TS嵌入,使得相似实例在文本坐标轴上的表示值相似,如等式3的第二项所示。

文本原型tp的建模功能是通过上一节描述的特征对比学习来实现的。特征矩阵m不再通过实例对比中使用的投影仪获得,而是通过原型映射e·tp→m获得。

可学习的提示嵌入即使使用 LLM 可以理解的嵌入表示来描述 TS,仍然需要指导 LLM 如何执行后续的 TS 任务。

如今,模板工程、思维链等硬提示方法直观、易懂,能够产生良好的效果。然而,它们的上下文在人类语义中是一致的。但正如我们提到的， TS 嵌入列表没有人类的语义,它更多的是关于模式序列。

因此,为了创建更一致的提示模式,我们训练了一个软提示,使 LLM 更容易理解输入（Lester、Al-Rfou 和 Constant 2021）。这些软提示是特定于任务的嵌入向量,通过LLM 输出和任务基本事实的损失进行学习,如公式 4 所示。软提示可以从均匀分布随机初始化,也可以从下游任务的文本嵌入初始化标签,或词汇peinit N (erandom/task/desc, σ)中最常见的单词。

Lpromp = Lreg/cls(concat(pe, e)) (4)

但同时,我们也相信有监督的微调方法可能提供更好的方法来提高TS 任务的准确性。然而,我们放弃的主要原因是,除了训练成本较高之外,LLM在微调后已经无法保证其理解人类语义文本的能力。这是因为 TS 嵌入的语义符合人类。

我们证明,经过训练的软提示不仅可以通过冻结LLM的语义理解能力来保护它,而且还可以达到与监督微调方法类似的效果:考虑一个条件生成任务,其中输入x是上下文,输出y是一个标记序列。

假设一个自回归 LLM pphi(y|x),参数为phi, z = [x; y]。预训练的 LLM 的推理是将hi作为zi及其左侧上下文中过去激活的函数来计算， Y = LMphi(zi , hi)。带有提示peθ的软提示调谐中过去的hi为方程5。从LLM到TS-LLM的微调为方程6。其变换表明软提示调谐近似等于微调。 peθ[i, :], 如果 i ∈ peidx LMphi(zi , hi)，否则

你好= (5)

算法1:TEST训练方法

1:对于历元中的e ，
= θfe = θfe－η▽θfe (Lins+Ltext) 更新编码器2: 3: // θfd
θfd － η▽θfd Lae θfp = θfp 更新解码器
4: － η▽θfp Lins 5:结束 更新投影仪
为6: for e
in epochs do 7: pe =
pe－η▽θpeLpromp // θfd = 更新提示
8: θfd － η '▽θfd Lreg 9: // 微调解码器
θfc = θfc － η▽θfc Lcls 10:结 更新分类器
束 for

最大Φ pψ(y ′ |x) = 最大值 对数pphi(z′ ≥ |h<i)
Φ i∈Yidx
= log pphi+Δ(zi + δzi |h<i)
i∈Yidx
≈ 对数pphi(zi |h<i) · log pΔ(δzi |h<i)
i∈Yidx i∈peidx
= 对数pψ(zi | 铁)
i∈Yidx
文本－TS 对齐
冷冻法学硕士
. log pΔ(δzi |h<i)
i∈peidx
提示peθ (6)

等式6还表明TS标记的投影空间应该优选地覆盖文本标记的完整嵌入空间集合。也就是说,我们需要选择这样的文本原型,它具有语义,更重要的是覆盖面广。因此,我们将LLM第一个嵌入层的表示空间的前k个主成分作为文本原型。

训练测试

TEST的核心是训练一个编码器fe和一个软提示pe。如算法 1 中所述,它们依赖于两个过程： fe通过对比学习进行训练,通过fp投影的特征对比和文本原型对齐的特征对比进行自我监督。 fd接受了预测任务的训练； pe由 LLM 的输出进行训练,并由类标签或真实值进行监督。对于分类任务,LLM 的 head fc需要进行训练。对于预测任务， fd需要进行微调。

使用LLM推断TS的过程如图2所示。在该框架中,文本数据被输入到LLM的嵌入层,而提示和TS嵌入则跳过该层。对于下游任务,TS分类任务需要在LLM fc之上额外增加一个头,就像实现NLP中的情感分类任务一样； TS 预测任务需要解码器fd与编码器一起训练。我们在算法中注释这些可选插件。

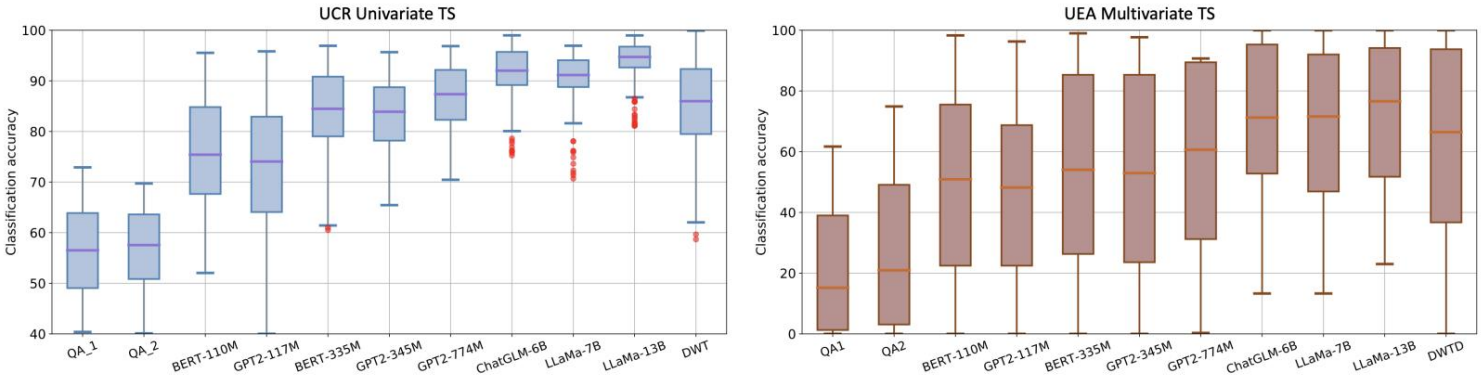

图 3:128 个 UCR 单变量 TS 数据集和 30 个 UEA 多变量 TS 数据集的方法分类精度

实验在本节中,我们在 UCR、UAE 和 TSER 数据集档案上实现 TS 分类和预测任务,以评估 TEST 和其他基线。

实现细节编码器模型是一个因果 TCN。我们为每个考虑的数据集存档选择一组超参数,无论下游任务如何。首先使用全连接层将每个时间步从多元 TS 的维度投影到大小为 64 的隐藏通道。

然后,有10层卷积块。每个卷积块都是 GELU、DilatedConv、BatchNorm、GELU、DilatedConv 的序列,每个块之间有跳跃连接。DilatedConvs 在卷积块的每层 i 中都有 2i 的膨胀,其中所有的内核大小均为 3,输入/输出通道大小为 64。最终的卷积块用于将隐藏通道映射到输出通道,其大小为与 LLM 的嵌入大小相同。

除非另有说明,否则每个编码器和软提示均使用 Adam优化器在 10 个配备 CUDA 11.3 的 NVIDIA Tesla V100-SXM2 GPU 上进行训练。

使用的 LLM 包括来自 Hugging Face1 的 Bert (Devlin et al. 2018)、GPT2 (Radford et al. 2019)、ChatGLM (Du et al. 2022) 和LLaMa2 (Touvron et al. 2023),如表 1 所列。 tokens的嵌入大小与LLM中文本token的嵌入大小相同。

模型	型号尺寸	嵌入尺寸
伯特	110M、335M	748、1024
GPT2	117M、345M、774M	768、1024、1280
聊天GLM 6B		4096
骆驼	7B、13B	4096

表1:使用的语言模型

1Hugging Face:https://huggingface.co
Bert:https://huggingface.co/bert-base-uncased;
GPT2:https://huggingface.co/gpt2;
ChatGLM:https://huggingface.co/THUDM/chatglm2-6b;
LLaMa:https://huggingface.co/meta-llama/Llama-2-7b。

分类

我们提供了UCR档案新迭代的所有 128 种不同单变量时间序列数据集的准确度分数（Dau 等人,2019）,这是 TS 分类任务的标准集。为了补充我们对仅包含单变量序列的 UCR 档案的评估,我们在多变量序列上评估我们的方法。这可以通过简单地改变所提出的编码器的第一卷积层的输入滤波器的数量来完成。我们在新发布的 UEA 档案（Bagnall 等人,2018)的所有 30 种不同的多元时间序列数据集上测试了我们的方法,这也是 TS 分类任务的标准集。

我们将我们的方法与使用或不使用提示的 LLM QA 方法（Xue 和 Salim 2023;Liu 等人 2023)以及 TS分类基线 DWT、DWTD（Bagnall 等人 2018)进行比较。

两个 QA 模板是： 1) [Q] 将给定的 [域]序列分类为 [类标签] 或 [类标签]:[数字序列]。 [A]; 2) [Q] 将具有[数值]平均值、[数值]方差和 [数值]采样率的序列分类为[类标签]或[类标签]。 [A]。同时,我们将多变量 TS 视为一系列单变量序列,并将它们依次填充到 QA 模板中。

我们不会将我们的方法与其他设计良好的基于深度学习的方法进行比较。因为我们不想让LLM成为专门针对TS的模式。我们的目标是保持其原有的语言能力,同时将其在 TS 任务上的性能从以前几乎不可能提高到与公认的基线（ UCR 的 ei DWT 和 UEA 的 DWTD)相当。

准确度所有方法的分类结果如图 3 所示。它们在 128 个 UCR 数据集和 30 个 UEA 数据集上的准确度得分分别用带有中位数和四分位数的箱线图表示。从结果中我们得出结论:TEST使得LLM的分类准确率显着提高。LLM 的原始分类性能通过两个 QA 结果来证明:它的准确性较低,几乎是随机猜测分类标签,

特别是对于多变量 TS。使用TEST后,所有模型中精度最低的 GPT2-117M可以IM-

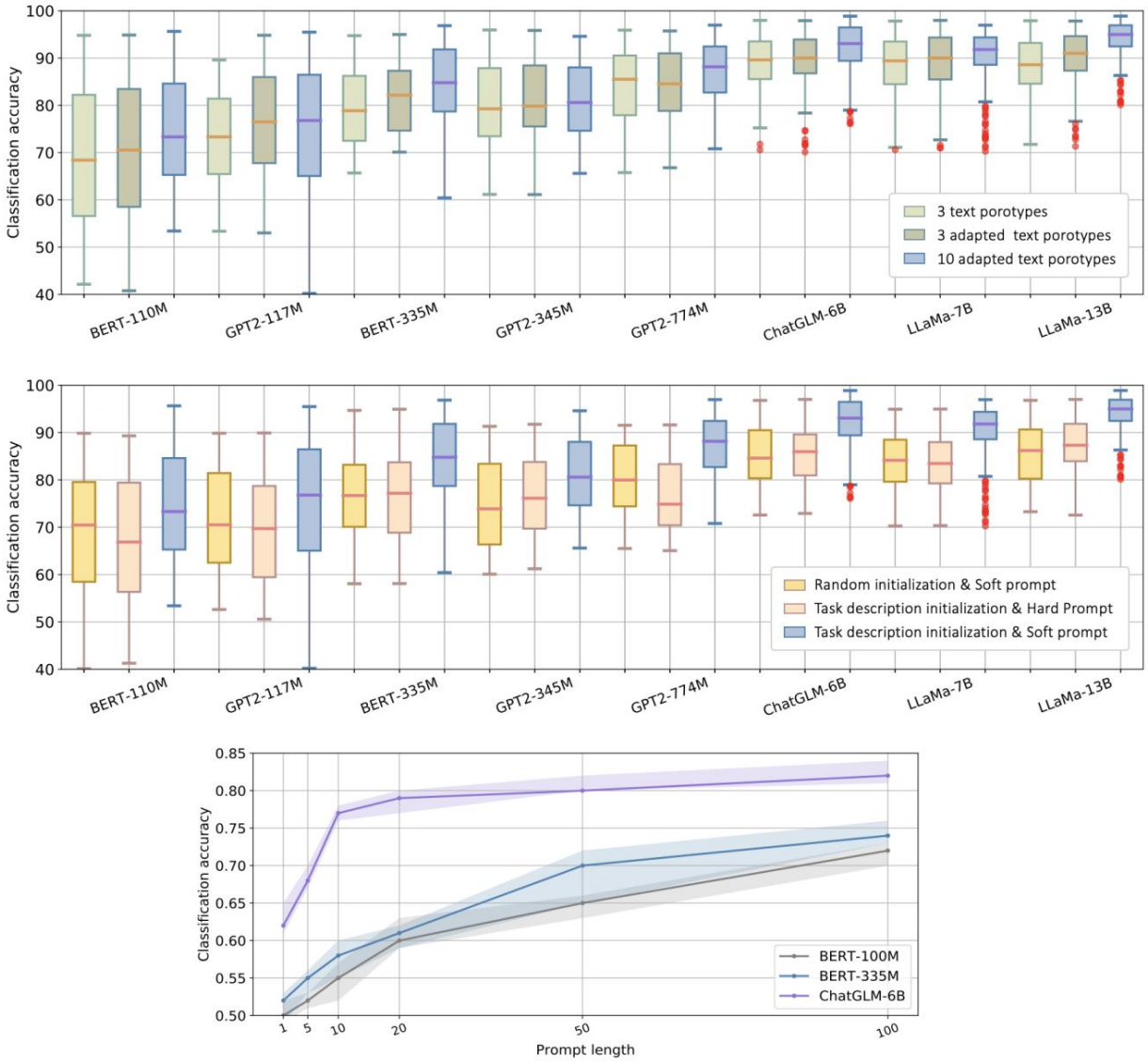

图 4:不同文本原型和提示的分类准确性

证明单变量 TS 的准确度至少为 18%,多变量 TS 的准确度至少为 25%。

TEST使大多数 LLM 与基线相当,甚至更好。当大小达到200M左右时,精度可以超过DWT;当大小达到700M左右时,精度可以超过DWTD。

模型的结构和尺寸会对结果产生影响。随着模型尺寸的增大,分类精度不断提高。在相同尺寸下,基于编码器的模型 Bert 和 ChatGLM 优于基于解码器的生成模型 GPT2 和 LLaMa。这也与编码器偏爱分类任务的做法一致（Du et al. 2022）。

文本原型的消融研究不同的文本原型会导致不同的结果。我们设置了三组文本原型:1)3个文本嵌入的文本原型值、形状和频率； 2)来自主成分分析的大致正交嵌入的 3 个自适应文本原型。 3)10个自适应文本原型。如图4 顶部所示,选择一个更准确地表示LLM 整个文本嵌入空间的原型组可以提高整体分类性能。公式 6也表明了这一点。同时,我们在图 1 中的前面示例中提供的第一个原型组的性能与第二个原型组的性能相当,这是因为这三个嵌入向量大致正交巧合的是。

提示的消融研究不同类型的提示会导致不同的结果。我们的方法训练了一个软提示,使 LLM 更容易理解输入。为了证明其有效性,我们将其与硬提示进行比较:将给定的[域]序列分类为[类标签]或

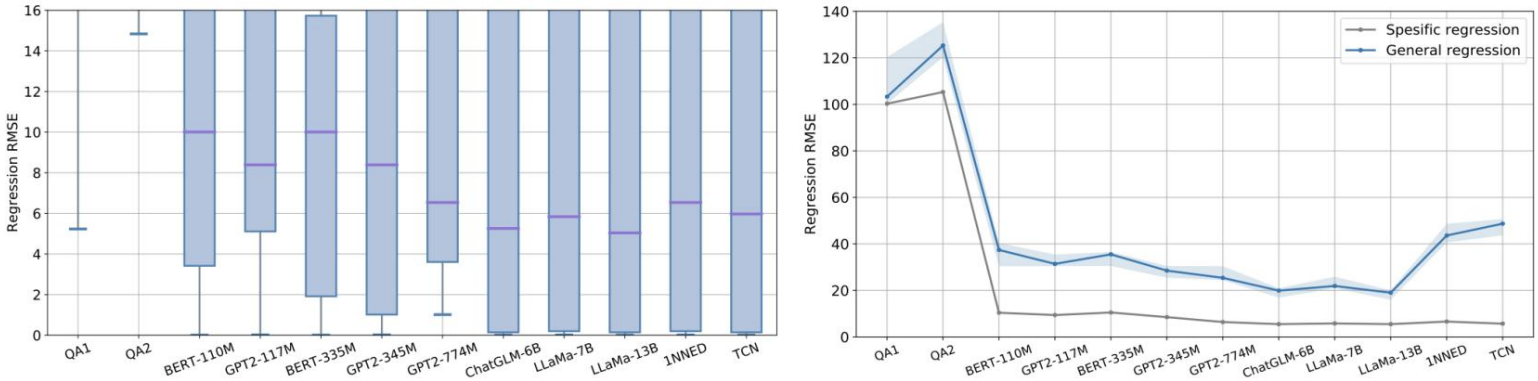

图 5:方法在 19 个 TSER TS 数据集上的预测精度

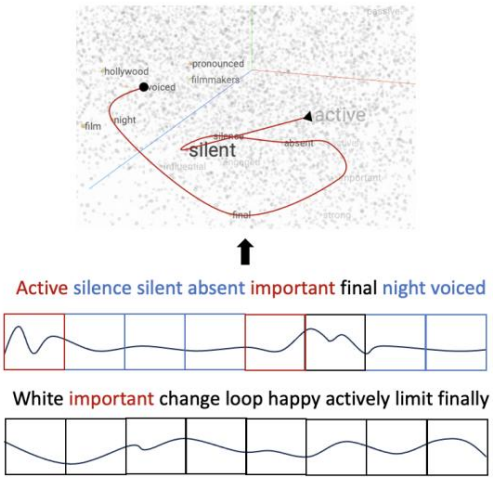

图 6:将 TS 嵌入与单词嵌入相匹配

[类标签]:[TS 嵌入]。如图4中所示,它们的准确度至少相差10%。同时,软提示的不同初始化也会导致不同的结果。

我们设置了两种提示初始化方法:1)均匀分布随机初始化； 2)根据给定序列分类的任务描述标记嵌入进行初始化。如图4中所示,非随机初始化的性能优于随机初始化。此外,不同长度的提示也会导致不同的结果。

如图 4 下方所示,当模型达到一定尺寸时,提示长度为 1 也能取得不错的效果,提示长度为 20 也能取得优异的效果。

案例研究我们使用最近邻法在冻结的 LLM 的词嵌入空间中查找 TS 标记匹配的文本。如图 6 的可视化所示,大多数单词都是有关情感的形容词和名词。

我们推测,通过提示,模型会将 TS 分类任务视为情感分类任务。因此,引入prompt就像引入LLM的捷径一样。

此外,匹配的单词就像一种用于TS 分割的文本 Shapelet,通过一系列模式。我们建议使用单词而不是数字来查找 TS 任务的模式。因为使用数字可能是一个序列推理任务,而现有的没有微调的法学硕士不擅长数学,但它们擅长作为模式机提取知识。

预测我们在

TSER 档案中提供所有 19 种不同时间序列数据集的准确度分数（Tan 等人,2021）。

我们将我们的方法与使用或不使用提示的 LLM QA 方法（Xue 和 Salim 2023;Liu 等人 2023)以及 TS预测基线 1NNED 和 TCN（Tan 等人 2021)进行比较。

两个 QA 模板是： 1) [Q] 预测给定 [域] 序列的下一个值:[数字序列]。 [A]; 2) [Q] 以[数值]的平均值、[数值]的方差、 [数值]的采样率预测序列的下一个值。 [A]。同时,我们将多变量 TS 视为一系列单变量序列,并将它们依次填充到 QA 模板中。

准确性所有方法的预测结果如图 5 所示。它们在 128 个 UCR 数据集和 30 个 UEA 数据集上的准确性分数分别用带有中位数和四分位数的箱线图表示。 TEST使得LLM的预测准确性显着提高并且与基线相当。原始LLM的QA模式无法实现预测任务,特别是如果TS数据被描述为均值、频率等（QA2）,则可能部分完成分类任务,但根本无法完成预测任务。

通用性我们将 19 个数据集融合为 1 个数据集,并在该融合数据集上测试该方法。与1NNED和TCN相比,基于LLM的模型具有更好的通用性。

讨论

- TS-for-LLM 的本质是:时间序列→标记→ TS 嵌入 ↔ ↔ 模式文本/词嵌入。

核心是将TS转换为LLM可以理解的模式序列。这个序列对应的文本语义可能会让我们感到困惑,但对法学硕士来说却很有意义。特别是在分类任务中,TS 模式是

由一些形容词表示,它们是语言模型分类任务的基础。这也反映出LLM在进行情感分类任务时更加注重带有情感的形容词。

· 在我们的方法中,影响任务有效性的主要因素是法学硕士的规模和类型、文本原型的选择以及提示的设计。

模型类型的影响与下游任务有关,其中双向结构有利于分类,生成的结构有利于预测。

更大的模型将使结果更准确。我们认为,这种现象的本质与用于预训练LLM的数据集有关。我们推测更多的训练数据使模型有更多的机会学习时间模式。根据图 5 右侧的观察,也许随着预训练中使用的数据集增多,原型的选择和提示的设计不再那么重要。为了确定其原因,我们打算在未来进行更多的实验,以研究语料库和时间序列之间更深层次的相关性。

· TS-for-LLM 可能不如训练小型任务导向模型那么高效和准确,但它可以丰富大型模型的能力并从其他角度探索其机制。

事实上,TS信息更像是位于文本和图像信息之间。它既表示顺序又表示形状。 TS 的抽象含义使得使用经典的基于对齐的多模态方法变得困难。我们的方法给人一种强制对齐操作的印象。我们希望这项工作能够激励未来的研究人员研究更好的融合方法。

结论

本文提出TEST,在TS-for-LLM的思想下,实现TS数据的实例级、特征级和文本原型对齐的嵌入方法。它可以激活LLM完成TS任务的能力,同时保持其原有的语言能力。分类和预测任务的实验表明,使用TEST， LLM 可以实现与 TS 基线相当的性能。未来的工作将测试其他TS 任务,例如异常检测,研究更多TS 和文本的对齐方法,甚至通过从头开始预训练 TS 基础大模型来实现 LLM-for-TS。

参考

阿劳克斯,J.;格雷夫,E.;库图里,M.;和 Joulin, A. 2019。多语言词嵌入的无监督超对齐。在国际学习表征会议上。

巴格纳尔,AJ;达乌,哈;莱恩斯,J.;弗林,M.;大,J。博斯特罗姆,A.;索瑟姆,P.;和 Keogh,EJ 2018。UEA多元时间序列分类档案,2018。CoRR， abs/1811.00075。

卡伦,M.;米斯拉,I.;迈拉尔,J.;戈亚尔,P.;博雅诺夫斯基,P.;和 Joulin, A. 2020。通过对比聚类分配进行视觉特征的无监督学习。在拉罗谢尔,H.;兰萨托,M.;哈德塞尔,R.;巴尔干,M.;和 Lin, H. 编辑,神经信息处理系统的进展。

陈,T。科恩布利斯,S.;诺鲁兹,M.;和辛顿,GE 2020。视觉表示对比学习的简单框架。国际机器学习会议记录,第 119 卷，1597-1607。

达乌,哈;巴格纳尔,A.;卡姆加尔,K.;叶,C.-CM;朱,Y。加尔加比,S.;拉塔纳马哈塔纳,加利福尼亚州;和基奥,E. 2019.UCR 时间序列档案。 IEEE/CAA自动化学杂志,6:1293–1305。

德夫林,J.;张,M。韭葱。;和图塔诺瓦,K. 2018。BERT:用于语言理解的深度双向变压器的预训练。 CoRR,abs/1810.04805。

杜,Z。钱,Y。刘X。丁,M。邱,J.;杨,Z。和 Tang, J. 2022。GLM:具有自回归空白填充的通用语言模型预训练。计算语言学协会年会记录,第 1 卷，320-335。

弗兰切斯基,J.;迪厄勒沃,A.;和 Jaggi, M. 2019。多元时间序列的无监督可扩展表示学习。神经信息处理系统的进展,4652–4663。

他,K。范,H.;吴,Y。谢S.;和吉尔希克,RB 2020。无监督视觉表示学习的动量对比。计算机视觉和模式识别,9726–9735 。

黄,Z。陈,J。张,J。 Shan, H. 2023。通过原型散射和正采样学习聚类表示。 IEEE 传输。模式肛门。马赫。情报， 45(6):7509–7524。

莱斯特,B.;阿尔-Rfou,R.； Constant, N. 2021。参数高效快速调整的规模力量。自然语言处理经验方法会议论文集， 3045-3059。

李,J。刘,C.;程,S。阿尔库奇,R.;和 Hong, S. 2023。冻结语言模型有助于心电图零样本学习。CoRR,abs/2303.12311。

李,Y。胡,P。刘JZ;彭,D.;周杰涛;和 Peng, X. 2021。对比聚类。 AAAI人工智能会议,8547–8555。

刘X。麦克达夫,D.;科瓦奇,G.； Galatzer-Levy,红外;阳光,JE;詹,J。Poh,M.;廖S.;阿奇尔,PD;和 Patel,SN 2023。大型语言模型是少数健康学习者。 CoRR,abs/2305.15525。

孟,Q。钱,H.;刘,Y。徐,Y。沉Z.;和崔,L. 2023.时间序列的无监督表示学习:回顾。 arXiv:2308.01578。

米尔昌达尼,S.;夏,F。佛罗伦萨,P.;伊希特,B.;德里斯,D.;阿里纳斯,MG;拉奥,K.;萨迪格,D.;和曾,A. 2023.作为通用模式机的大型语言模型。 arXiv:2307.04721。

雷德福,A.;吴,J。孩子,R。栾,D.;阿莫代伊,D.;和Sutskever, I. 2019。语言模型是无监督的多任务学习者。开放人工智能。

谭,CW;伯格梅尔,C.;珀蒂让,F.;和韦伯,GI 2021。
时间序列外在回归。数据挖掘和知识发现,1-29。

图夫龙,H.;马丁,L.;斯通,K.;阿尔伯特,P.;阿尔马海里,A.;和等人。
2023. Llama 2:开放基础和微调聊天模型。 arXiv:2307.09288。

王,T。 Isola, P. 2020。通过超球面的对齐和均匀性理解对比表征学习。国际机器学习会议录,第 119 卷,9929–9939。

吴,G。刘,C.;萨胡,D.;库马尔,A.;和海,SCH
2022.CoST:时间序列预测的解开季节趋势表示的对比学习。在国际学习代表会议上。

薛,H.;和 Salim,FD 2023。PromptCast:一种新的基于提示的时间序列预测学习范式。 arXiv:2210.08964。

岳,Z。王,Y。段,J。杨,T。黄,C.;童,Y。和 Xu, B. 2022。TS2Vec:迈向时间序列的通用表示。 AAAI 人工智能会议， 8980-8987。

郑X;陈X.;舒尔奇,M.;莫莱萨,A.;阿拉姆,A.;和 Krauthammer, M.
2023。SimTS:重新思考时间序列预测的对比表示学习。 CoRR，
abs/2303.18205。