复现 SAMformer: Unlocking the Potential of Transformers in Time Series Forecasting with Sharpness-Aware Minimization and Channel-Wise Attention

摘要

Transformer 架构在自然语言处理和计算机视觉领域实现了显著性能提升，但在多变量长期预测任务上，其表现仍不及简单的线性基准。为探究这一现象，我们首先针对一个线性预测的玩具问题进行了研究，结果表明尽管 Transformer 具有极高的表达能力，却无法收敛至其真实解。我们还发现，Transformer 的注意力机制是导致其泛化能力不足的关键因素。基于这一发现，我们提出了一种轻量级的浅层 Transformer 模型，结合了尖锐度感知优化以摆脱局部最优解。实验证明，这一改进在所有主流的真实世界多变量时间序列数据集上均有效。特别是，SAMformer 不仅超越了现有先进方法，而且在参数数量上大幅减少，与大规模模型 MOIRAI 性能相当。相关代码可在https://github.com/romilbert/samformer获取。

关键词：通道注意力机制；时间序列预测；RevIN

1 引言

多元时间序列预测是一个经典的学习问题，旨在根据历史信息预测未来趋势。在现实世界的许多应用中，如医疗数据、电力消耗、气温、股票价格等领域，时间序列预测都具有重要意义。然而，尽管 Transformer 架构在自然语言处理和计算机视觉中取得了突破性的性能，但在多元时间序列预测中，当前最先进的方法是基于 MLP 的模型，而 Transformer 方法的表现却不如预期。近期研究发现，线性网络在预测任务中甚至可以与 Transformer 相媲美或更优，这一现象引发了对 Transformer 在时间序列预测中应用的思考。

作者以Zeng et al. [2023] 等人的研究结果为起点，通过构建一个模拟时间序列预测设置的玩具回归问题，发现即使将 Transformer 架构调整为解决简单的线性预测问题，它仍然泛化能力较差且收敛到局部极小值。进一步研究表明，Transformer 的注意力机制是导致这一问题的主要原因。

通过分析 Transformer 在时间序列预测任务中的当前缺陷，并设计适当的训练策略来解决这些问题，作者提出的 SAMformer 模型在多元长期时间序列预测中表现出了优越的性能。该研究不仅有助于提升 Transformer 在时间序列预测中的表现，还为未来相关研究提供了新

的思路和方向。此外，SAMformer 在计算效率、对预测 horizons 的鲁棒性以及对不同初始化的稳定性等方面具有优势，具有重要的实际应用价值。

2 相关工作

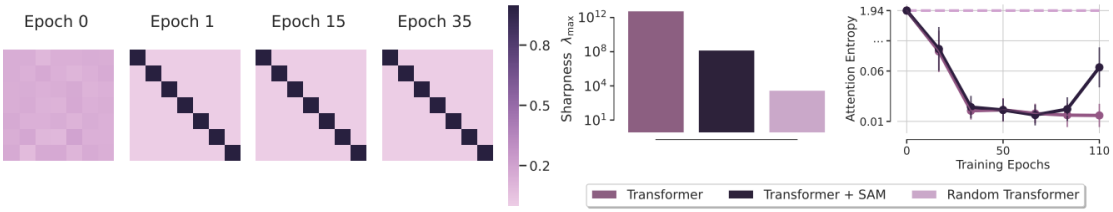
2.1 Transformer 在时间序列预测中的应用

许多工作尝试提出时间序列特定的 Transformer 架构，以利用其捕捉时间交互的能力。但当前在多元时间序列预测中，基于 MLP 的模型表现更优，且线性网络在预测任务中有时能与 Transformer 表现相当，这使 Transformer 在该任务中的实用性受到质疑。

2.2 Transformer 的训练问题

1. 训练不稳定：在计算机视觉和自然语言处理中，发现注意力矩阵可能存在秩崩溃问题。已有一些方法被提出以解决这些问题，但在时间序列预测中，Transformer 的训练稳定性问题尚未得到充分解决，特别是在缺乏大规模数据的情况下。如图1所示：

2. 损失景观尖锐：研究表明 Transformer 的损失景观比其他残差架构更尖锐，这可能解释了其训练不稳定和在小规模数据集上表现不佳的原因。如1所示：



(a) Attention matrices of `Transformer` along the training.

(b) Sharpness at the end of the training, Entropy collapse.

图 1. Transformer 模型损耗景观的线性回归分析如下：(a) 从训练的第一个周期开始，Transformer 的注意力矩阵即表现出一致性偏好。(B，左侧)Transformer 相较于融合了 SAM 策略的 Transformer+SAM，收敛至更为尖锐的最小值，其最大特征值 $\lambda_{\max}$ 较大（达到 100×10^4），而随机 Transformer 则展现出较为平滑的损耗分布。。(B，右侧)Transformer 在训练过程中出现熵崩溃现象，这进一步验证了其损失景观的高度清晰性。

2.3 解决方法的研究

1. 现有方法：一些研究通过计算损失函数 Hessian 的最大特征值 $\lambda_{\max}$ 或衡量注意力矩阵的熵来量化 Transformer 的尖锐度，并提出了一些解决方法，如使用 sharpness-aware minimization（SAM）框架或通过谱归一化和额外学习标量进行重参数化（aReparam）。

   - Foret et al. [2021] 最近提出的锐度感知最小化框架。这种方法通过替换训练目标 $\mathcal{L}_{\text{train}}$ 来实现，新的训练目标 $\mathcal{L}_{\text{train}}^{\text{SAM}}(\omega)$ 定义为：

$$\mathcal{L}_{\text{train}}^{\text{SAM}}(\omega) = \max_{\|\epsilon\| < \rho} \mathcal{L}_{\text{train}}(\omega + \epsilon)$$

     其中，$\rho > 0$ 是一个超参数，ω 是模型的参数。

- 第二种方法涉及使用谱归一化和额外的学习标量来重新参数化所有权重矩阵，这种方法由Zhai et al. [2023] 提出，称为 oReparam。具体来说，每个权重矩阵 W 被替换为：

$$W_c = \frac{\gamma}{\|W\|_2} W$$

其中，γ 是一个可学习的参数，初始值为 1。

2. 本文方法：作者提出了 SAMformer 模型，通过结合可逆实例归一化（RevIN）Kim et al. [2022] 和使用 SAM 进行优化，以解决 Transformer 在时间序列预测中的问题。实验表明，SAMformer 在多元长期时间序列预测中优于现有的预测基线，包括最大的现有时间序列预测基础模型 MOIRAI，且具有更高的通用性和鲁棒性。

综上所述，作者通过分析 Transformer 在时间序列预测任务中的现有问题，并提出 SAMformer 模型及相应的训练策略，以提高其在该任务中的性能。

3 本文方法

模型架构

SAMformer 基于 Transformer 架构，并对其进行了改进，以适应时间序列预测的需求。主要改进包括引入了 Sharpness-Aware Minimization (SAM) 和 Channel-Wise Attention (CWA) 机制。架构图2：

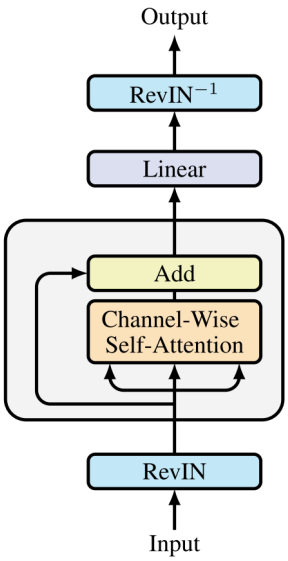


图 2. SAMformer

Transformer 架构通常包括自注意力机制（Self-Attention）和位置编码（Positional Encoding），形式化表示为：

$$\text{Output} = \text{Self-Attention}(Q, K, V) + \text{Positional Encoding}$$

其中，Q, K, 和 V 分别代表查询（Query）、键（Key）和值（Value）矩阵。

Sharpness-Aware Minimization (SAM)

SAM 是一种优化策略，它通过在损失函数中加入一个关于模型参数的平滑度正则项，来帮助模型逃离局部最优解，从而提高泛化能力。Sharpness 损失可以用以下公式表示：

$$\text{Sharpness Loss} = \frac{1}{2} \sum_{\theta} \nabla_{\theta}^2 \text{Loss}(\theta)$$

其中，θ 表示模型参数，∇_{θ}^2 表示对参数的二阶导数。

Channel-Wise Attention (CWA)

CWA 是对 Transformer 注意力机制的改进，它允许模型在不同的时间步长和通道之间动态地分配注意力权重，增强了模型在时间序列数据中的时序和频率特征的表达能力。CWA 的注意力权重计算如下：

$$\text{CWA} = \text{Softmax}\left(\text{Weight Matrix}\right) \cdot \text{Feature Maps}$$

损失函数

基础损失函数

论文中使用的基础损失函数通常是均方误差（MSE）或其他常用的回归损失函数，用于衡量预测值与真实值之间的差异：

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

其中，y_i 是真实值，$\hat{y}_i$ 是预测值，N 是样本数量。

Sharpness 损失

SAM 引入了一个额外的损失项，即 Sharpness 损失，它基于模型参数的二阶导数，鼓励模型参数空间的平滑性。

算法

训练算法

使用带有 Sharpness 损失的正则化损失函数进行训练：

$$\text{Total Loss} = \text{Base Loss} + \alpha \cdot \text{Sharpness Loss}$$

其中，α 是 Sharpness 损失的系数。

采用梯度下降或其变体来最小化整体损失，包括基础损失和 Sharpness 损失。在训练过程中，应用 Channel-Wise Attention 来动态调整不同通道的注意力权重。

预测算法

使用训练好的 SAMformer 模型进行时间序列的预测。根据模型的输入序列，利用 Transformer 架构和 CWA 机制生成未来的时间序列预测：

$$\hat{y}_{t+\Delta t} = \text{SAMformer}(x_t, \Delta t)$$

其中，$\hat{y}_{t+\Delta t}$ 是未来时间 $t + \Delta t$ 的预测值，x_t 是当前时间 t 的输入序列，Δt 是预测的时间步长。

4 实验

4.1 与已有开源代码对比

该论文提供了源码,在实验上主要对比了 SAMformer 超越了当前的最先进方法(TSMixer)以及将 SAM 应用到 TSMixer 上（注：这是作者原有的实验我先跑完这些确定 SAM+ 通道注意力在 Transformer 领域是否有提升，结果是显著的），而我的工作不仅跑了项目的原有实验，而且在常见的 7 个多元长期预测数据集上（比如说：ETT 系列、traffic、Electricity 等数据集）又尝试了 SAM+Transformer 和 Transformer 的对比以及 SAM+Random Transformer 和 Random Transformer 的对比，效果显示加了 SAM 后的固定注意力 Radom Transformer 在多个数据集上取得领先，具体细节在表1查看。

4.2 数据集

我们在 8 个公开可用的真实世界时间序列数据集上进行了实验，这些数据集被广泛用于多变量长期预测。4 个电力变压器温度数据集 ETTm1、ETTm2、ETTh1 和 ETTh2 包含了从 2016 年 7 月至 2018 年 7 月电力变压器收集的时间序列。我们将这些 4 个数据集统称为 ETT。Electricity 包含了从 2012 年到 2014 年 321 个客户的电力消耗时间序列。Exchange Rate 包含了从 1990 年到 2016 年 8 个国家之间每日汇率的时间序列。Traffic（加利福尼亚交通部，2021 年）包含了从 2015 年 1 月至 2016 年 12 月由 862 个传感器捕获的道路占用率时间序列。最后但同样重要的是，Weather（马克斯 · 普朗克研究所，2021 年）包含了由 21 个气象指标在 2020 年记录的气象信息时间序列。应该注意的是，电力、交通和天气都是大规模数据集。

数据集可以在这里下载Download datasets。表2总结了我们在实验中使用的数据集的特性。

5 总结与展望

这篇论文提出了 SAMformer，这是一种基于 Transformer 架构的新型时间序列预测模型。SAMformer 通过引入 Sharpness-Aware Minimization (SAM) 和 Channel-Wise Attention (CWA) 机制来解锁 Transformer 在时间序列预测中的潜力。

表 1. Transformer 和 Ranom Transformer 在加与不加 SAM 之间的性能比较，用于具有不同视野的多变量长期预测 H。我们显示了 epoch10 次后获得的平均测试 MSE。较低的结果表示较好的预测。最佳结果以粗体显示。

Dataset	**H**	with SAM		**Without SAM**	
		transformer	Random Transformer	transformer	Random Transformer
ETTh1	96	0.404	**0.384**	0.434	0.385
	192	0.433	0.453	0.449	**0.423**
	336	0.469	**0.457**	0.495	0.480
	720	0.535	**0.522**	0.599	0.560
ETTh2	96	0.342	**0.303**	0.392	0.310
	192	0.475	0.416	0.482	**0.394**
	336	0.480	**0.427**	0.539	0.466
	720	0.944	0.779	0.924	**0.768**
ETTm1	96	0.314	**0.306**	0.328	0.312
	192	0.346	**0.341**	0.354	0.360
	336	0.380	0.382	**0.376**	**0.376**
	720	**0.434**	0.452	0.445	0.438
ETTm2	96	0.183	0.182	0.196	**0.179**
	192	0.270	**0.239**	0.255	0.248
	336	0.317	0.306	**0.289**	0.309
	720	0.414	0.422	**0.401**	0.439
electricity	96	0.201	0.162	**0.161**	0.162
	192	0.200	**0.172**	0.173	**0.172**
	336	0.206	0.184	**0.182**	0.184
	720	0.228	0.213	**0.210**	0.213
exchange	96	**0.115**	0.120	0.135	0.126
	192	0.254	0.241	0.300	**0.216**
	336	0.438	0.471	1.049	**0.373**
	720	**0.493**	1.073	2.574	0.836
traffic	96	1.100	**0.511**	3.232	0.517
	192	0.908	**0.502**	2.216	0.504
	336	0.770	**0.494**	1.948	0.497
	720	0.700	**0.510**	1.763	0.509
Weather	96	0.187	0.178	**0.174**	0.178
	192	0.224	**0.218**	0.219	**0.218**
	336	0.266	0.262	0.261	0.261
	720	0.323	0.323	0.323	**0.320**

表 2. 各种大小和维度的多变量时间序列数据集的特征

Dataset	ETTh1/ETTh2	ETTm1/ETTm2	Electricity	Exchange	Traffic	Weather
# features	7	7	321	8	862	21
# time steps	17420	69680	26304	7588	17544	52696
Granularity	1 hour	15 minutes	1 hour	1 day	1 hour	10 minutes

5.1 SAMformer 的核心创新

Sharpness-Aware Minimization (SAM): SAM 通过在损失函数中加入一个关于模型参数的平滑度正则项，帮助模型逃离局部最优解，从而提高泛化能力。这有助于 SAMformer 在长期预测任务中表现更好。查看图3以便更好的理解。

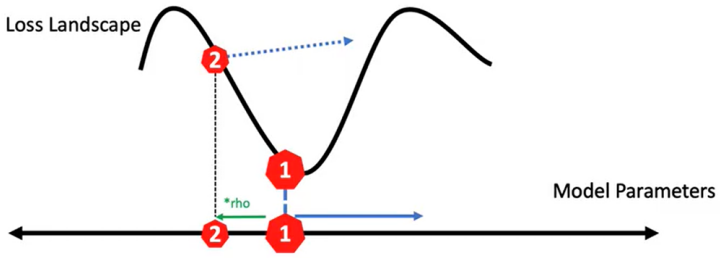


图 3. 从损失函数的几何特性入手，考虑最小值附近的平坦度，使用在第二个位置计算的梯度来更新原始位置的参数，迫使模型移动到新区域。

Channel-Wise Attention (CWA): CWA 允许模型在不同的时间步长和通道之间动态地分配注意力权重，增强了模型在时间序列数据中的时序和频率特征的表达能力。

5.2 实验结果

论文报告了 SAMformer 在各种公开可用的真实世界时间序列数据集上的实验结果，包括 ETT 数据集、电力消耗数据集、汇率数据集、交通数据集和天气数据集。实验结果表明，SAMformer 在多变量长期预测任务中取得了显著的性能提升。接着，我们提供了具有平均绝对误差 MAE 不同模型的比较，并证实加 SAM 之后有模型的确有极大的提升。具体细节在表3查看。

5.3 不足与展望：

尽管 SAMformer 在时间序列预测方面取得了显著进展，但仍存在一些不足之处。例如，SAMformer 的模型复杂度较高，计算成本较大，这可能会限制其在实际应用中的可扩展性。此外，CWA 机制在捕获时间序列数据的复杂特征方面仍有改进空间。

5.4 未来研究方向：

- **模型压缩与优化：** 研究如何降低 SAMformer 的模型复杂度，提高计算效率，使其能够更好地应用于实际场景。

表 3. Transformer 和 Ranom Transformer 在加与不加 SAM 之间的性能比较，用于具有不同视野的多变量长期预测 H。我们显示了 epoch10 次后获得的平均绝对误差 MAE。较低的结果表示较好的预测。最佳结果以粗体显示。

Dataset	**H**	**with SAM**		**without SAM**	
		Transformer	Random Transformer	Transformer	Random Transformer
ETTh1	96	0.426	**0.408**	0.457	0.409
	192	0.442	0.463	0.459	**0.435**
	336	0.469	**0.459**	0.489	0.481
	720	0.537	**0.527**	0.576	0.554
ETTh2	96	0.401	**0.369**	0.438	0.372
	192	0.484	0.445	0.476	**0.423**
	336	0.490	**0.453**	0.526	0.473
	720	0.689	0.637	0.684	**0.634**
ETTm1	96	0.358	**0.351**	0.370	0.356
	192	0.378	**0.374**	0.384	0.395
	336	0.402	0.402	0.396	**0.398**
	720	**0.435**	0.455	0.444	0.440
ETTm2	96	0.281	0.280	0.302	**0.278**
	192	0.349	**0.322**	0.332	0.328
	336	0.380	0.368	**0.350**	0.373
	720	0..435	0.449	**0.428**	0.452
Electricity	96	0.308	0.267	**0.266**	0.267
	192	0.306	**0.275**	0.276	0.276
	336	0.311	0.288	0**.287**	**0.287**
	720	0.328	0.313	**0.312**	0.313
Exchange	96	0.260	**0.255**	0.270	0.263
	192	0.391	0.365	0.408	**0.345**
	336	0.510	0.510	0.755	**0.461**
	720	0.785	0.762	1.207	**0.678**
Traffic	96	0.716	**0.391**	1.176	0.395
	192	0.633	**0.380**	0.991	0.382
	336	0.558	**0.371**	0.941	0.374
	720	0.514	0.370	0.889	**0.369**
Weather	96	0.245	0.235	**0.234**	0.237
	192	0.279	**0.272**	0.278	**0.272**
	336	0.313	0.310	0.310	**0.308**
	720	0.361	0.362	0.362	**0.357**

- **CWA 机制的改进:** 探索更有效的 CWA 机制，以更好地捕获时间序列数据的复杂特征，提高预测精度。
- **多任务学习和自适应学习:** 研究如何将 SAMformer 应用于多任务学习场景，以及如何使其能够根据不同的任务和数据集自动调整模型结构和参数。
- **与其他时间序列模型的比较:** 将 SAMformer 与其他时间序列预测模型进行比较，分析其优缺点，并探索如何将 SAMformer 与其他模型相结合，以进一步提高预测性能。

5.5 总结

SAMformer 为时间序列预测提供了一种新的思路和方法，通过 SAM 和 CWA 机制，在多变量长期预测任务中取得了显著的性能提升。尽管仍存在一些不足，但 SAMformer 为未来时间序列预测研究提供了有价值的方向和启示。

参考文献

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40770–40803. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.