2024 ICML CCF-A

SAMformer

Unlocking the Potential of Trans**former**s in Time Series Forecasting with **S**harpness-**A**ware **M**inimization and Channel-Wise Attention

Transformer-based

时间序列预测

锐度感知最小化(SAM)

解决训练不稳定性

ICML2024

华为巴黎研究中心

2024 ICML CCF-A

SAMformer

Unlocking the Potential of Trans**formers** in Time Series Forecasting with **S**harpness-**A**ware **M**inimization and Channel-Wise Attention

24.6.4

Presented by Yyyq

01 问题描述

- 目前的SOTA模型：**TSMixer（完全线性模型）**
- 目前的Transformer-based工作主要集中在：
  - Reduce the cost：降低注意力机制的N^2成本
  - Decompose TS：解耦时间序列捕获潜在模式
- 仍然没有解决Transformer的**训练不稳定性**（倾向于过拟合，泛化能力差）
  - 陷入局部最小值，且很难在随后的迭代中退出

➡ **模型收敛到尖锐的局部最小值**

[1] Chen S A, Li C L, Arik S O, et al. TSMixer: An All-MLP Architecture for Time Series Forecast-ing[J]. Transactions on Machine Learning Research, 2023.

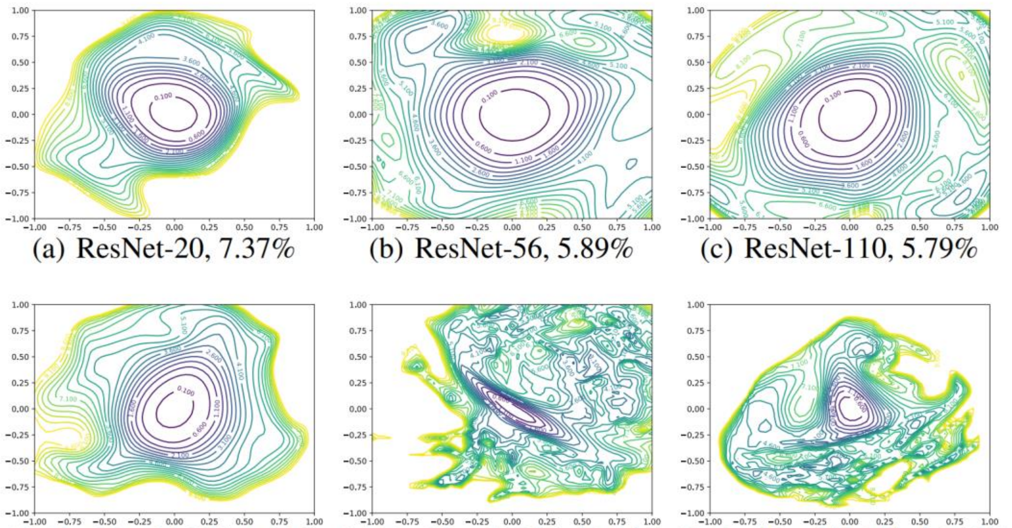
02 背景补充：The Sharpness of the **Loss** Landscape

- 我们训练神经网络的目的是：最小化损失函数
- 损失函数它是一个“高度非凸”的函数
  - 定义域内存在多个局部极小（大）值，而不是仅有一个全局最小（大）值
  - 我们无法简单地沿梯度下降找到全局最小值
- 这种“非凸”来自于：
  - 网络深度的增加，参数数量的增加（参数复杂性），非线性激活函数（ReLU，tanh）
- 由“非凸”带来的挑战：
  - 优化算法可能会陷入局部极小值，而不是找到全局最优解
- 我们目前的哪些方法是在努力克服这个问题：
  - 随机梯度下降SGD，自适应调整学习率Adam，正则化dropout

02 背景补充：The Sharpness of the **Loss** Landscape

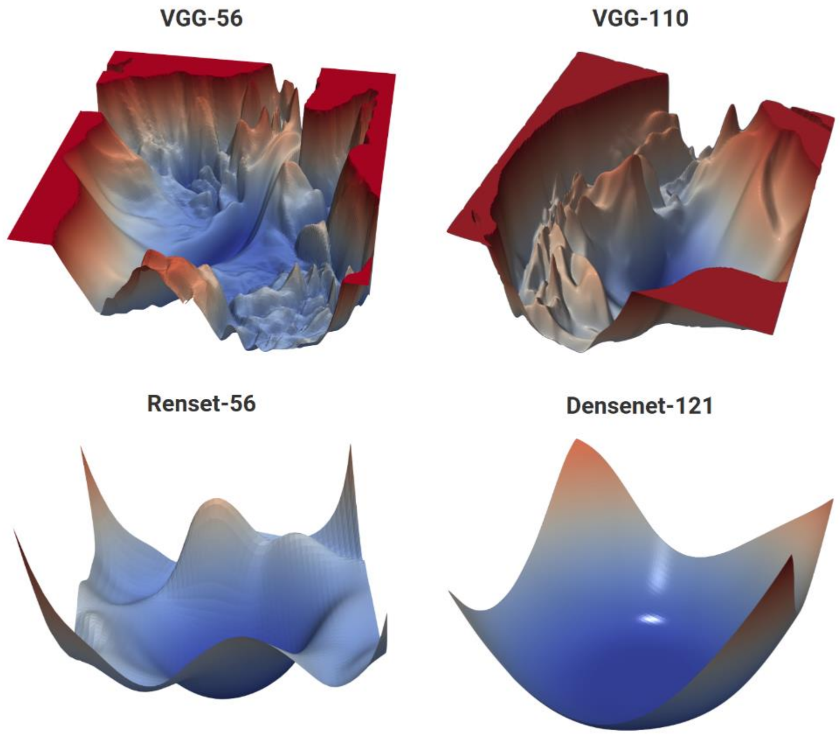
➢ 例如：ResNet残差连接

- 产生更容易训练的损失函数，且避免过拟合问题
- 产生更加平坦的损失曲面，提高泛化能力



(a) ResNet-20, 7.37% (b) ResNet-56, 5.89% (c) ResNet-110, 5.79%

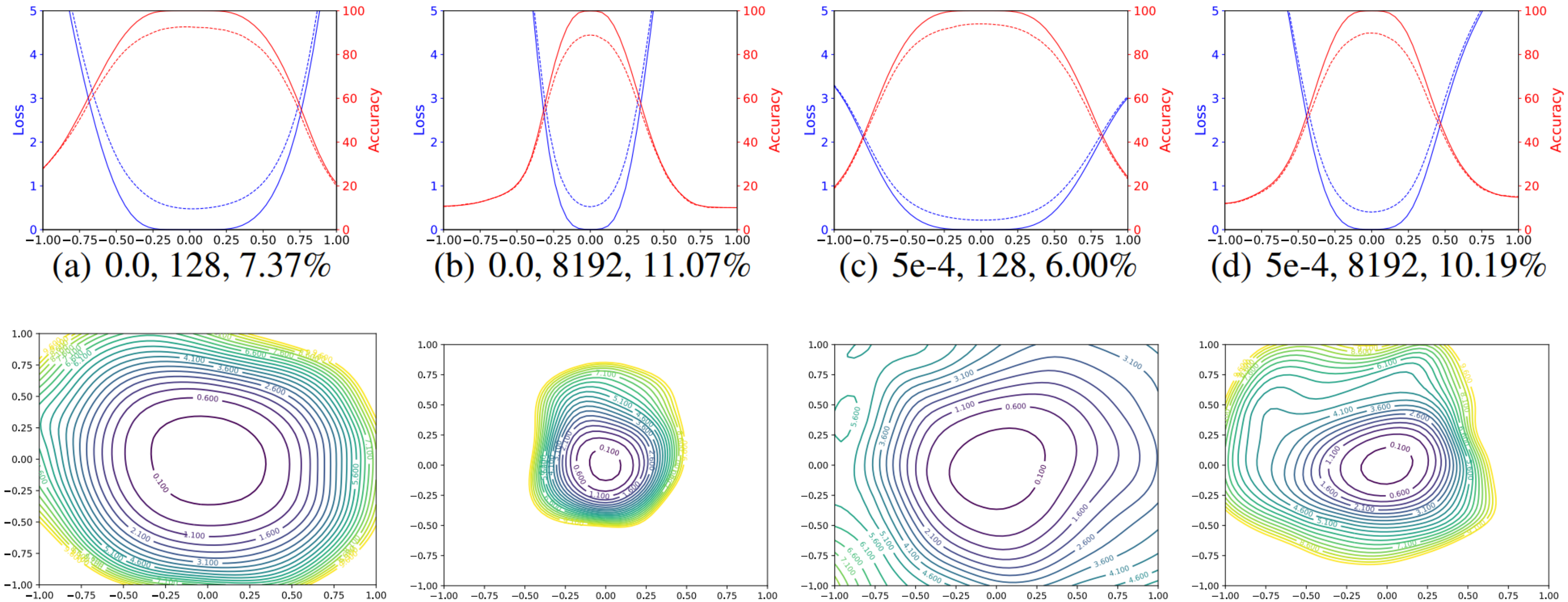
(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%



[1]马里兰大学NIPS2018：Visualizing the Loss Landscape of Neural Nets（可视化神经网络的损失情况）

02 背景补充：The **Sharpness** of the **Loss** Landscape

➢ **局部最小化解的锐度/平坦度(“sharp” vs “flat”)与泛化能力之间的关系**



(a) 0.0, 128, 7.37%　(b) 0.0, 8192, 11.07%　(c) 5e-4, 128, 6.00%　(d) 5e-4, 8192, 10.19%

(e) 0.0, 128, 7.37%　(f) 0.0, 8192, 11.07%　(g) 5e-4, 128, 6.00%　(h) 5e-4, 8192, 10.19%

[1]马里兰大学NIPS2018：Visualizing the Loss Landscape of Neural Nets（可视化神经网络的损失情况）

背景补充：The **Sharpness** of the **Loss** Landscape

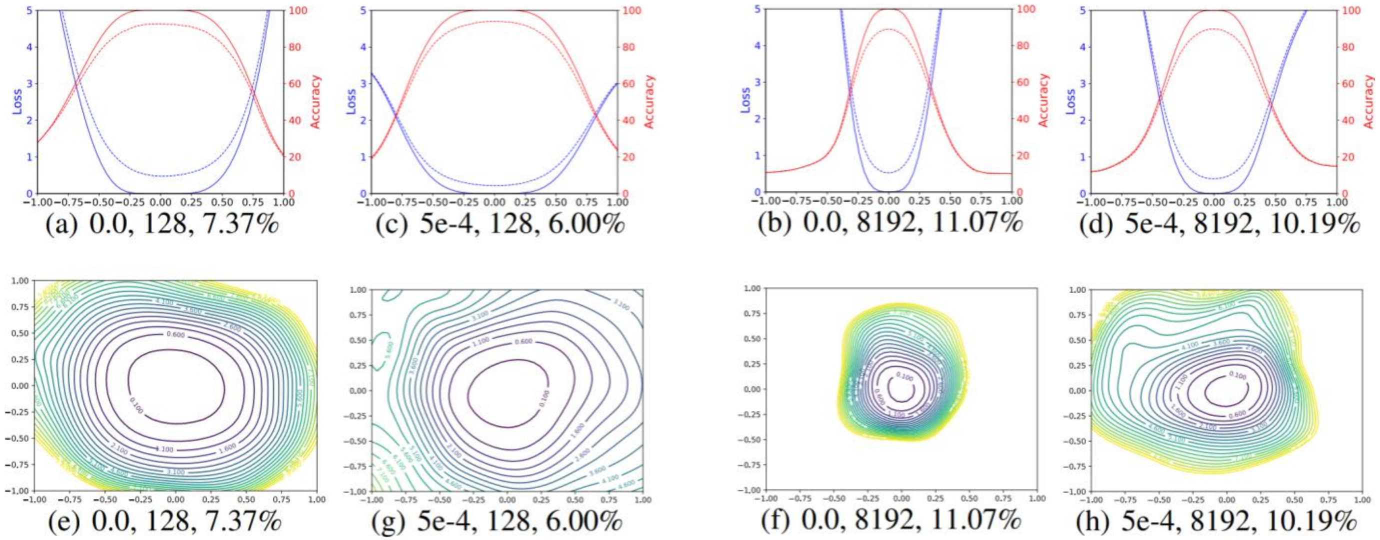
➢ **局部最小化解的锐度/平坦度(“sharp” vs “flat”)与泛化能力之间的关系**

- 平坦的损失曲面表示模型对参数变化不敏感，这通常与更好的泛化性能相关
- 批量大小（batchsize）和权重衰减（weight decay）对模型训练结果的影响

  →看测试误差→看损失曲面的形状（sharpness）→模型的泛化能力

加入权重衰减

相对平坦



Batchsize = 128 Batchsize = 8192

[1]马里兰大学NIPS2018：Visualizing the Loss Landscape of Neural Nets（可视化神经网络的损失情况）

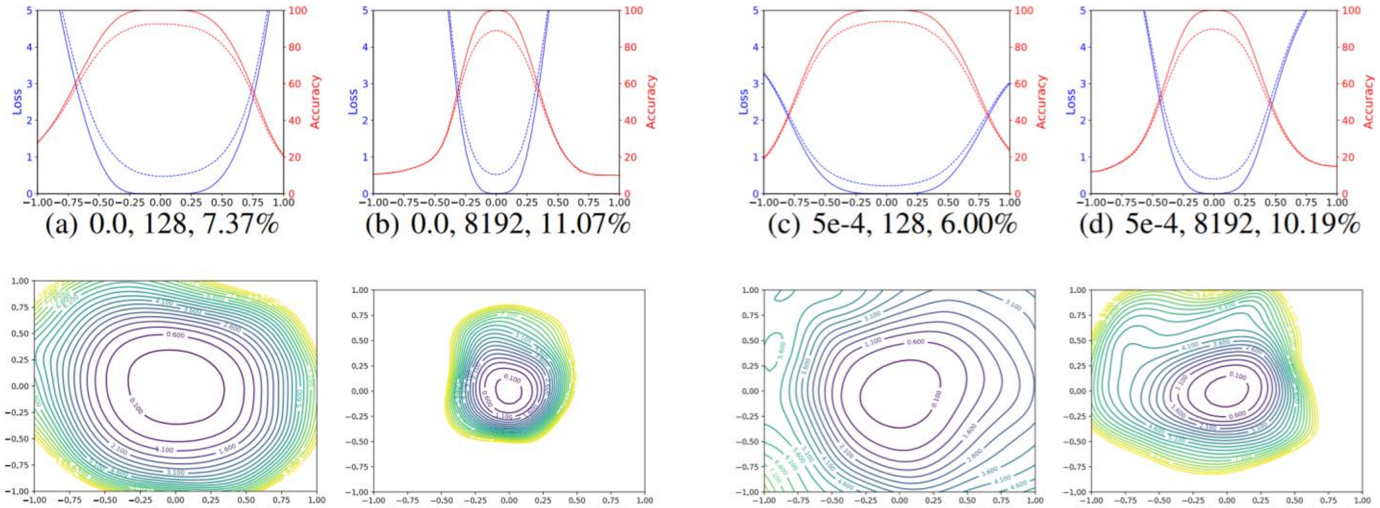
02 背景补充：The **Sharpness** of the **Loss** Landscape

- **局部最小化解的锐度/平坦度(“sharp” vs “flat”)与泛化能力之间的关系**
  - 平坦的损失曲面表示模型对参数变化不敏感，这通常与更好的泛化性能相关
  - 批量大小（batchsize）和权重衰减（weight decay）对模型训练结果的影响

    →看测试误差→看损失曲面的形状（sharpness）→模型的泛化能力

小批量

相对平坦



(a) 0.0, 128, 7.37%　(b) 0.0, 8192, 11.07%　(c) 5e-4, 128, 6.00%　(d) 5e-4, 8192, 10.19%

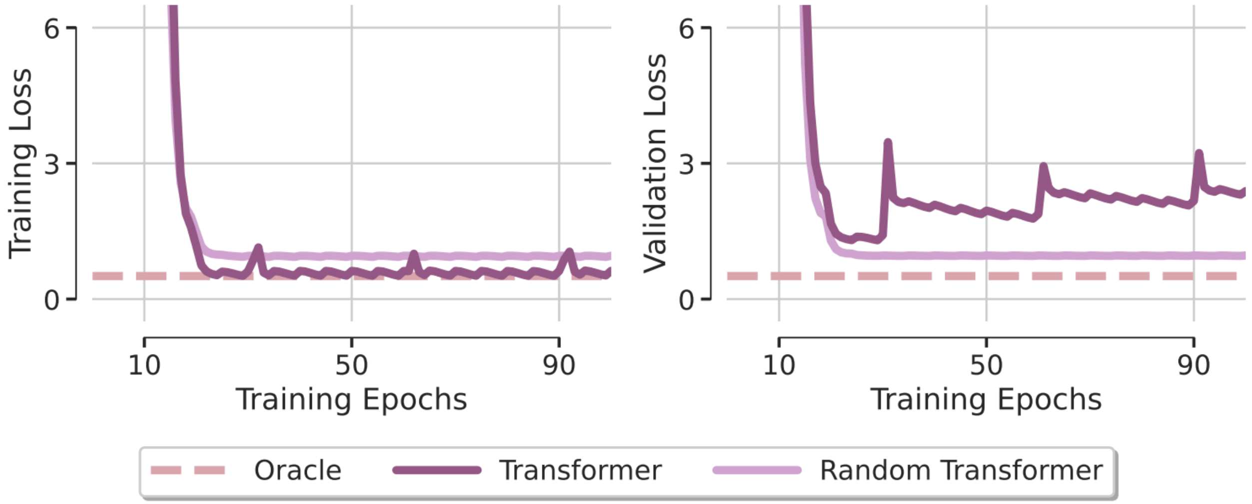
(e) 0.0, 128, 7.37%　(f) 0.0, 8192, 11.07%　(g) 5e-4, 128, 6.00%　(h) 5e-4, 8192, 10.19%

weight decay = 0.0　　weight decay = 5e-4

[1]马里兰大学NIPS2018：Visualizing the Loss Landscape of Neural Nets（可视化神经网络的损失情况）

03 动机举例：A Toy Regression Problem

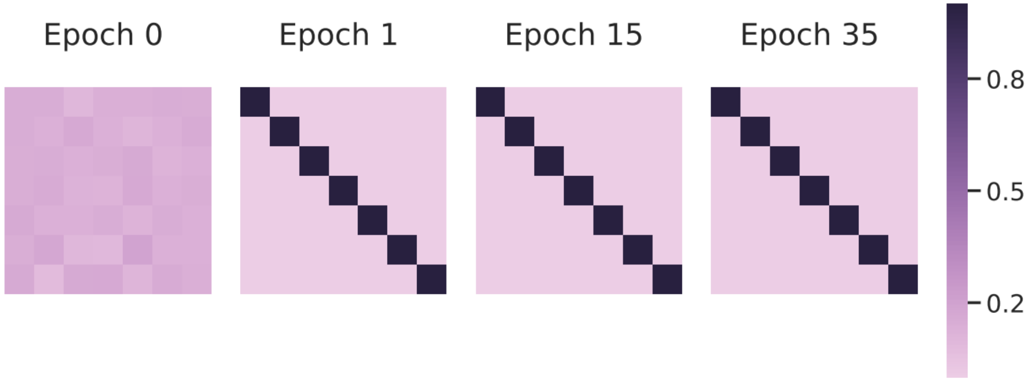
- 定义一个生成模型
  - 模拟时序预测： $\mathbf{Y} = \mathbf{X}\mathbf{W}_{\text{toy}} + \varepsilon.$
- 简化的transformer编码器
  - 注意力机制+残差连接+线性层： $f(\mathbf{X}) = [\mathbf{X} + \underline{\mathbf{A}(\mathbf{X})}\underline{\mathbf{X}\mathbf{W}_V}\mathbf{W}_O]\mathbf{W},$
  - 基于通道的注意力矩阵： $\mathbf{A}(\mathbf{X}) = \text{softmax}\left(\frac{\underline{\mathbf{X}\mathbf{W}_Q}\underline{\mathbf{W}_K^\top\mathbf{X}^\top}}{\sqrt{d_{\text{m}}}}\right) \in \boxed{\mathbb{R}^{D\times D}}$

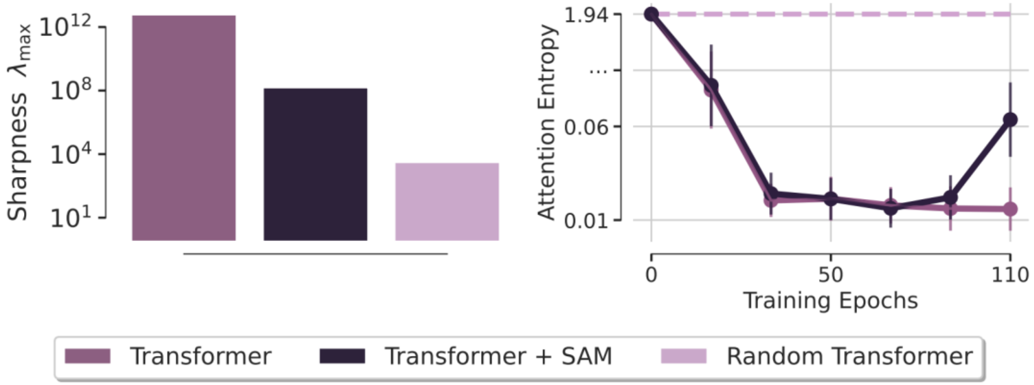


- Oracle：利用最小二乘法计算W，得到理论最优解
- Random Transformer：只有W被优化，attention权重$W_Q W_K$ W_V W_O固定，使得所考虑的Transformer像一个线性模型。
- Transformer：通过注意力机制和残差连接来拟合

04 Transformer's Loss Landscape

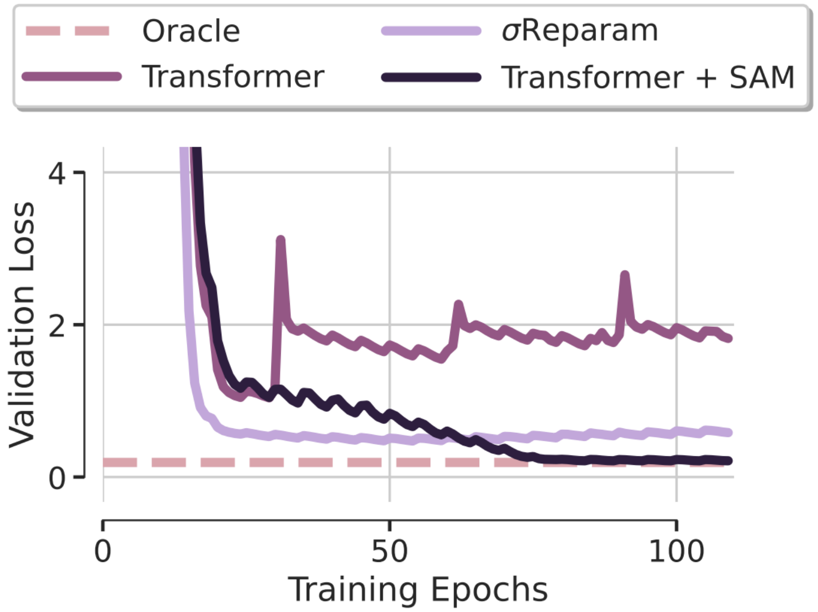
- ➢ 注意力矩阵
  - 优化的 transformer 可能陷入次优局部最小值
  - the sharpness of the loss landscape of the transformer
- ➢ 损失函数曲线
  - 固定注意力的Transformer(Random Transformer)的sharpness比收敛到单位矩阵的Transformer(Transformer)低几个数量级
  - 注意力矩阵的熵随着训练轮次的增加急剧下降（熵崩溃→过拟合、训练不稳定）



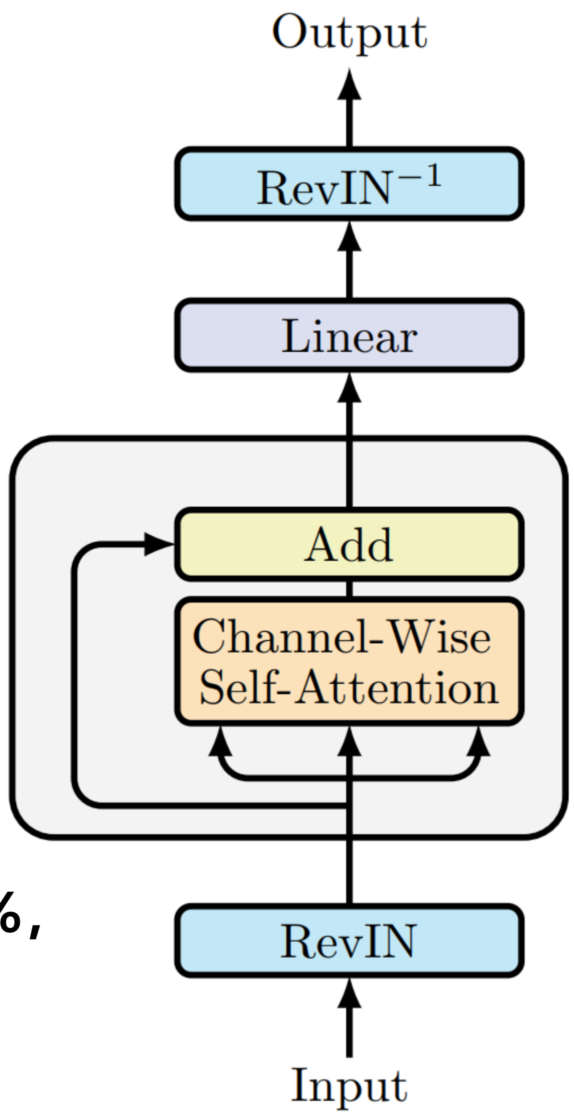


05 SAMformer

- A single layer with one attention head
  - RevIN：可逆实例归一化
  - Channel-Wise Self-Attention
- 使用SAM来优化模型，以使其收敛到更平坦的局部最小值。



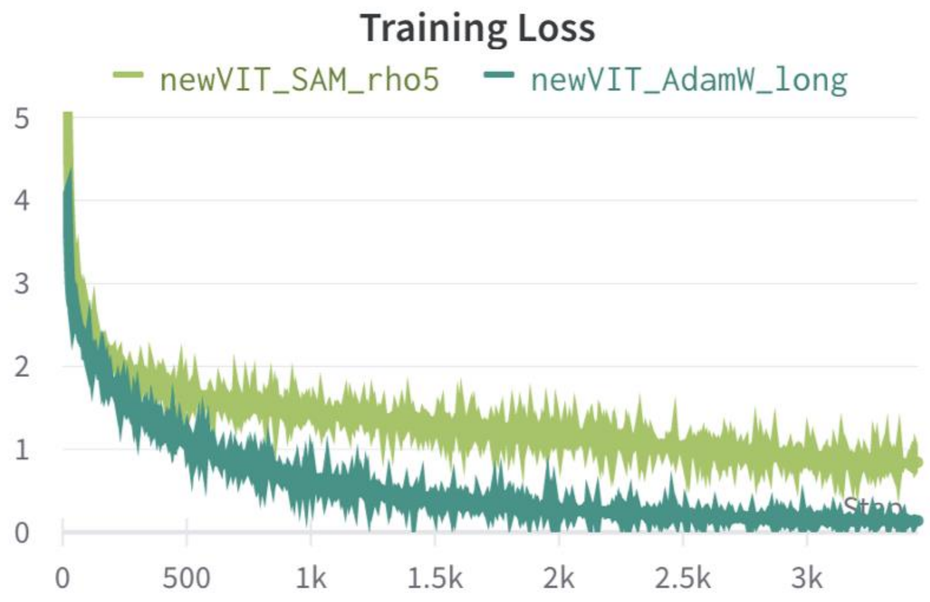
1. **具有更好的泛化能力**
2. **相比于TSMixer提高了14.33%，同时参数减少了约4倍**

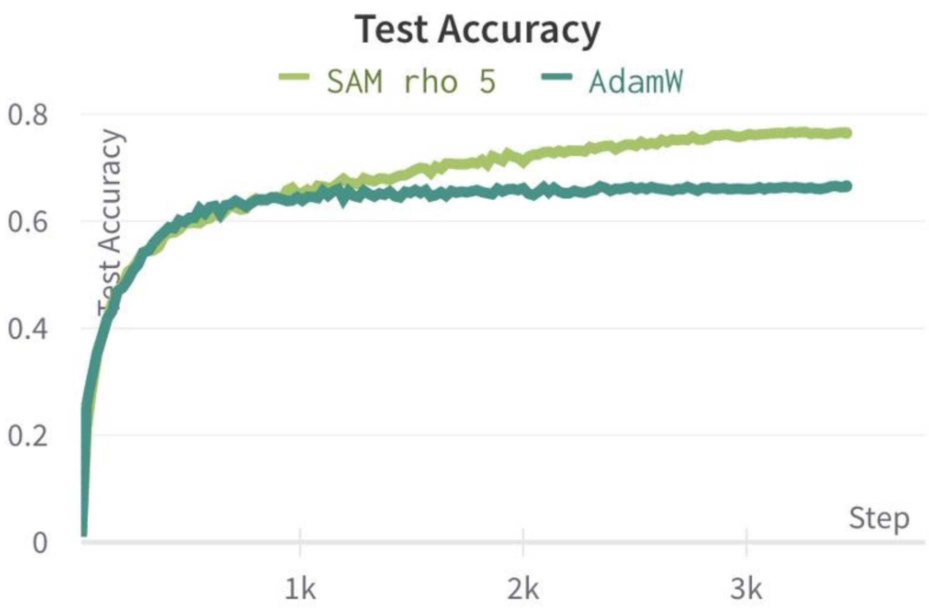


05 **SAMformer：Sharpness-Aware Minimization**

➢ 同时最小化损失值和损失曲面的锐度 → 提高模型的泛化能力

“现代深度神经网络的损失函数往往是非凸的，具有多个局部甚至全局极小值，这些极小值可能在训练集上表现相似，但在测试集上的表现差异很大”





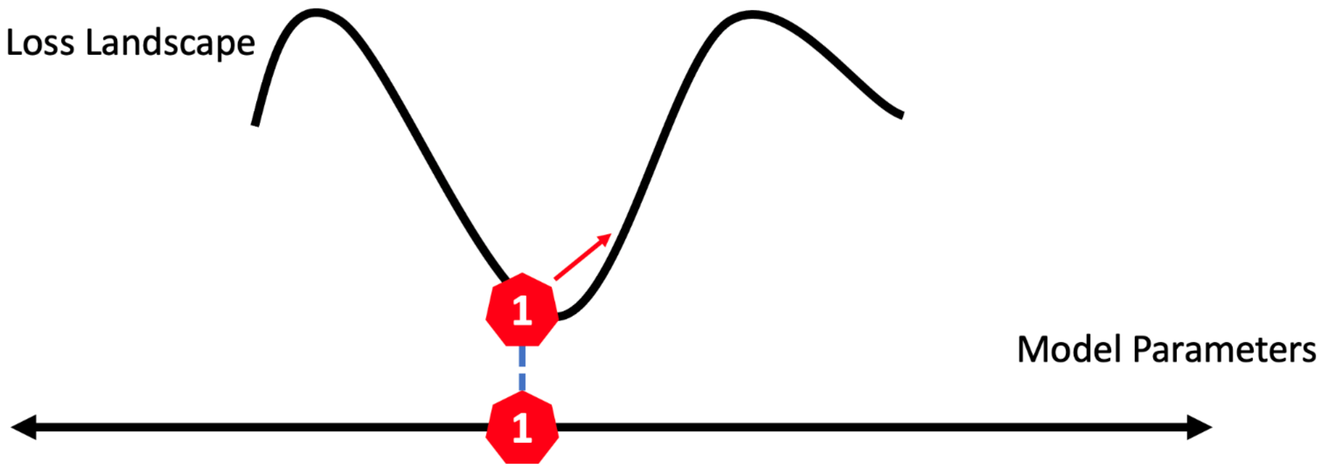
[1]谷歌研究院ICLR2021：Sharpness-aware minimization for efficiently improving generalization（SAM）

05

SAMformer：Sharpness-Aware Minimization

➢ 从损失函数的几何特性入手，考虑最小值附近的平坦度

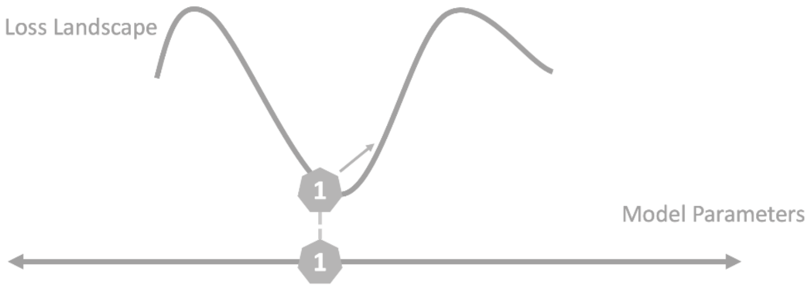
- 当模型进入一个尖锐的最小值时，通常的梯度更新会导致模型在该最小值附近振荡。

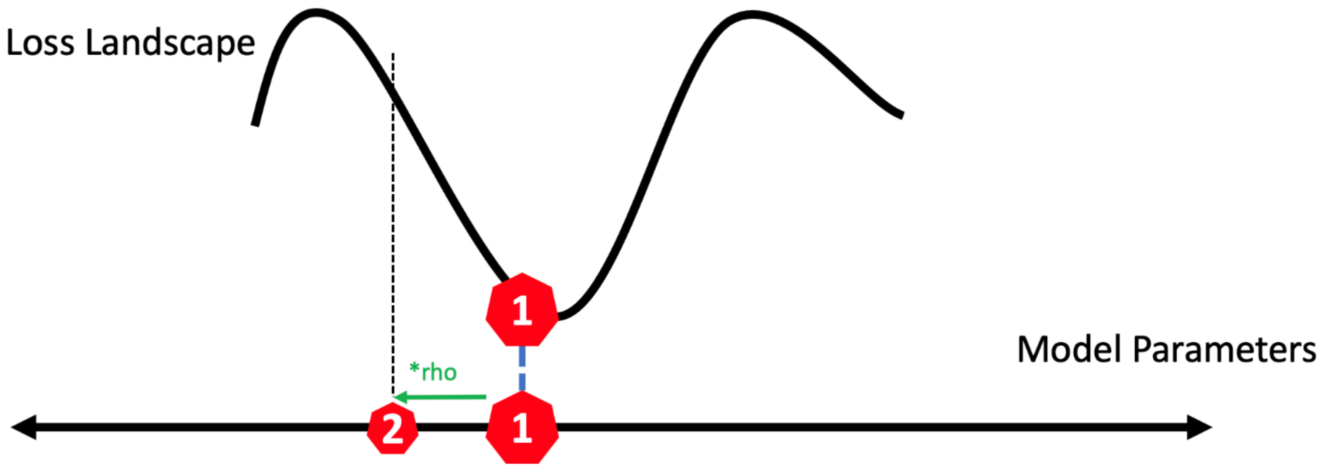


[1]谷歌研究院ICLR2021：Sharpness-aware minimization for efficiently improving generalization（SAM）

05 SAMformer：Sharpness-Aware Minimization

- 从损失函数的几何特性入手，考虑最小值附近的平坦度
  - SAM通过计算梯度后，采取相反的方向移动，这个反向移动由一个因子 ρ 缩放。



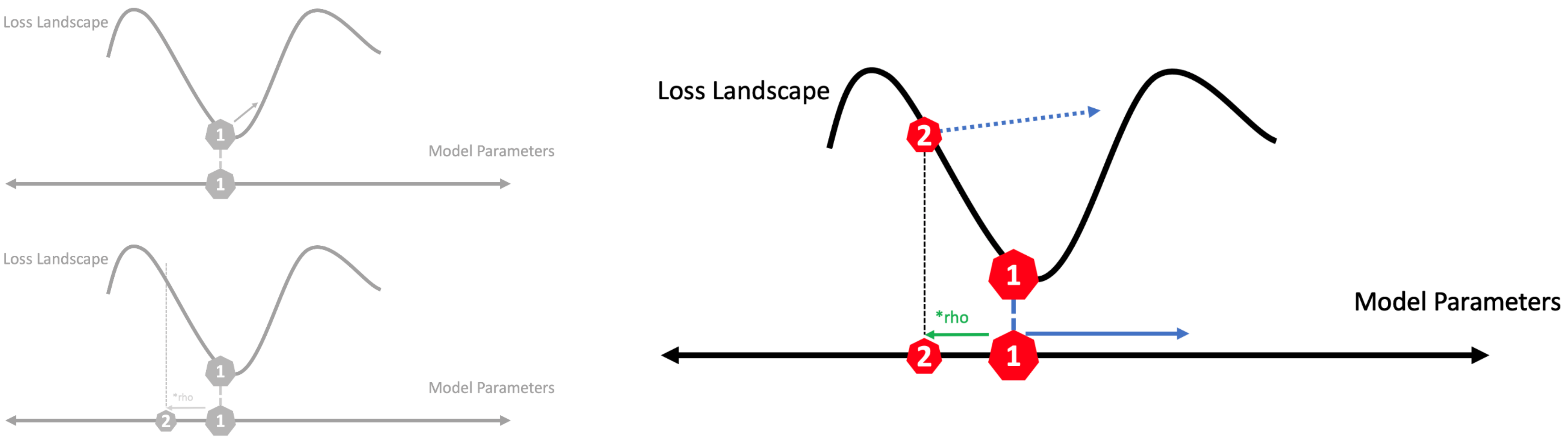


[1]谷歌研究院ICLR2021：Sharpness-aware minimization for efficiently improving generalization（SAM）

05 SAMformer：Sharpness-Aware Minimization

➢ 从损失函数的几何特性入手，考虑最小值附近的平坦度

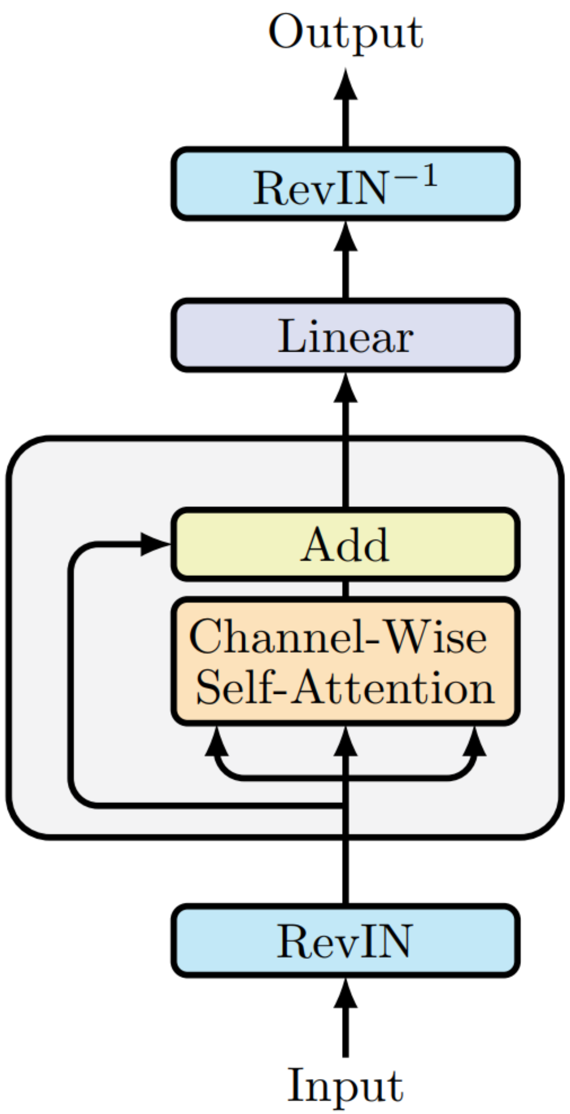
• 使用在第二个位置计算的梯度来更新原始位置的参数，迫使模型移动到新区域。



[1]谷歌研究院ICLR2021：Sharpness-aware minimization for efficiently improving generalization（SAM）

05 SAMformer

- A single layer with one attention head
  - RevIN：可逆实例归一化
  - Channel-Wise Self-Attention
- 使用SAM来优化模型，以使其收敛到更平坦的局部最小值。
  - 它并不是直接替代传统优化器，而是与它们协同工作，以实现更好的训练效果。



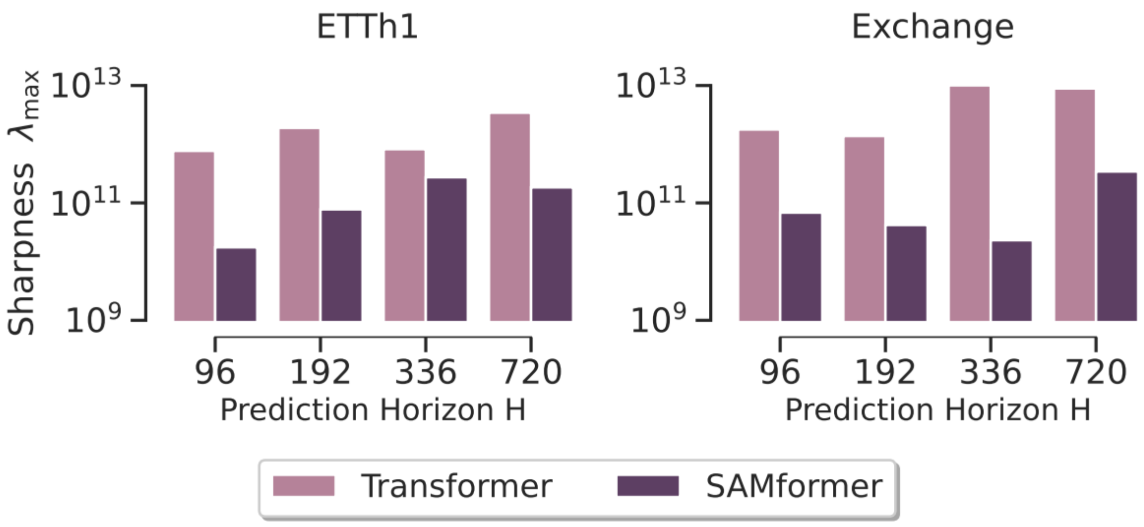
06

实验1：对比实验

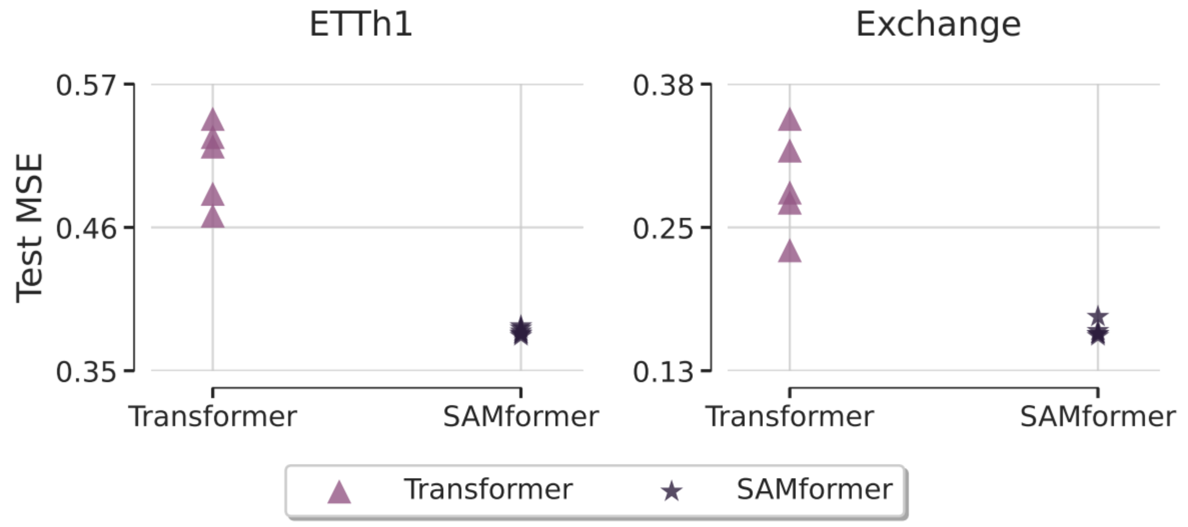
输入长度L = 512，预测H∈{96,192,336,720}

Dataset	H	with SAM		without SAM						
		SAMformer	TSMixer	Transformer	TSMixer	In*	Auto*	FED*	Pyra†	LogTrans†
ETTh1	96	0.381±0.003	0.388±0.001	0.509±0.031	0.398±0.001	0.941	0.435	**0.376**	0.664	0.878
	192	**0.409**±0.002	0.421±0.002	0.535±0.043	0.426±0.003	1.007	0.456	0.423	0.790	1.037
	336	**0.423**±0.001	0.430±0.002	0.570±0.016	0.435±0.003	1.038	0.486	0.444	0.891	1.238
	720	**0.427**±0.002	0.440±0.005	0.601±0.036	0.498±0.076	1.144	0.515	0.469	0.963	1.135
ETTh2	96	**0.295**±0.002	0.305±0.007	0.396±0.017	0.308±0.003	1.549	0.332	0.332	0.645	2.116
	192	**0.340**±0.002	0.350±0.002	0.413±0.010	0.352±0.004	3.792	0.426	0.407	0.788	4.315
	336	**0.350**±0.000	0.360±0.002	0.414±0.002	0.360±0.002	4.215	0.477	0.400	0.907	1.124
	720	**0.391**±0.001	0.402±0.002	0.424±0.009	0.409±0.006	3.656	0.453	0.412	0.963	3.188
ETTm1	96	0.329±0.001	0.327±0.002	0.384±0.022	0.336±0.004	0.626	0.510	**0.326**	0.543	0.600
	192	**0.353**±0.006	0.356±0.004	0.400±0.026	0.362±0.006	0.725	0.514	0.365	0.557	0.837
	336	**0.382**±0.001	0.387±0.004	0.461±0.017	0.391±0.003	1.005	0.510	0.392	0.754	1.124
	720	**0.429**±0.000	0.441±0.002	0.463±0.046	0.450±0.006	1.133	0.527	0.446	0.908	1.153
ETTm2	96	0.181±0.005	0.190±0.003	0.200±0.036	0.211±0.014	0.355	0.205	**0.180**	0.435	0.768
	192	**0.233**±0.002	0.250±0.002	0.273±0.013	0.252±0.005	0.595	0.278	0.252	0.730	0.989
	336	**0.285**±0.001	0.301±0.003	0.310±0.022	0.303±0.004	1.270	0.343	0.324	1.201	1.334
	720	**0.375**±0.001	0.389±0.002	0.426±0.025	0.390±0.003	3.001	0.414	0.410	3.625	3.048
Electricity	96	**0.155**±0.002	0.171±0.001	0.182±0.006	0.173±0.004	0.304	0.196	0.186	0.386	0.258
	192	**0.168**±0.001	0.191±0.010	0.202±0.041	0.204±0.027	0.327	0.211	0.197	0.386	0.266
	336	**0.183**±0.000	0.198±0.006	0.212±0.017	0.217±0.018	0.333	0.214	0.213	0.378	0.280
	720	**0.219**±0.000	0.230±0.005	0.238±0.016	0.242±0.015	0.351	0.236	0.233	0.376	0.283
Exchange	96	0.161±0.007	0.233±0.016	0.292±0.045	0.343±0.082	0.847	0.197	**0.139**	-	0.968
	192	**0.246**±0.009	0.342±0.031	0.372±0.035	0.342±0.031	1.204	0.300	0.256	-	1.040
	336	**0.368**±0.006	0.474±0.014	0.494±0.033	0.484±0.062	1.672	0.509	0.426	-	1.659
	720	**1.003**±0.018	1.078±0.179	1.323±0.192	1.204±0.028	2.478	1.447	1.090	-	1.941
Traffic	96	**0.407**±0.001	0.409±0.016	0.420±0.041	0.409±0.016	0.733	0.597	0.576	2.085	0.684
	192	**0.415**±0.005	0.433±0.009	0.441±0.039	0.637±0.444	0.777	0.607	0.610	0.867	0.685
	336	**0.421**±0.001	0.424±0.000	0.501±0.154	0.747±0.277	0.776	0.623	0.608	0.869	0.734
	720	**0.456**±0.003	0.488±0.028	0.468±0.021	0.688±0.287	0.827	0.639	0.621	0.881	0.717
Weather	96	0.197±0.001	**0.189**±0.003	0.227±0.012	0.214±0.004	0.354	0.249	0.238	0.896	0.458
	192	0.235±0.000	**0.228**±0.004	0.256±0.018	0.231±0.003	0.419	0.325	0.275	0.622	0.658
	336	0.276±0.001	**0.271**±0.001	0.278±0.001	0.279±0.007	0.583	0.351	0.339	0.739	0.797
	720	0.334±0.000	**0.331**±0.001	0.353±0.002	0.343±0.024	0.916	0.415	0.389	1.004	0.869
Overall MSE improvement			5.25%	16.96%	14.33%	72.20%	22.65%	12.36%	61.88%	70.88%

06 实验2



(a) Sharpness of `SAMformer` and `Transformer`.



(b) Performance across runs of `SAMformer` and `Transformer`.

更平滑的损失曲面

对随机初始化更具有鲁棒性

06 实验3：计算效率

Dataset	$H = 96$		$H = 192$		$H = 336$		$H = 720$		Total
	SAMformer	TSMixer	SAMformer	TSMixer	SAMformer	TSMixer	SAMformer	TSMixer	
ETT	50272	124142	99520	173390	173392	247262	369904	444254	-
Exchange	50272	349344	99520	398592	173392	472464	369904	669456	-
Weather	50272	121908	99520	171156	173392	245028	369904	442020	-
Electricity	50272	280676	99520	329924	173392	403796	369904	600788	-
Traffic	50272	793424	99520	842672	173392	916544	369904	1113536	-
Avg. Ratio	6.64		3.85		2.64		1.77		**3.73**

① 更少的超参数：a single layer with one attention head（仅用了缩放点积）

② 超参数具有更强的通用性：在不同数据集上的应用更加简便

③ 更好的注意力机制：SAMformer在特征之间强烈促进自相关

06

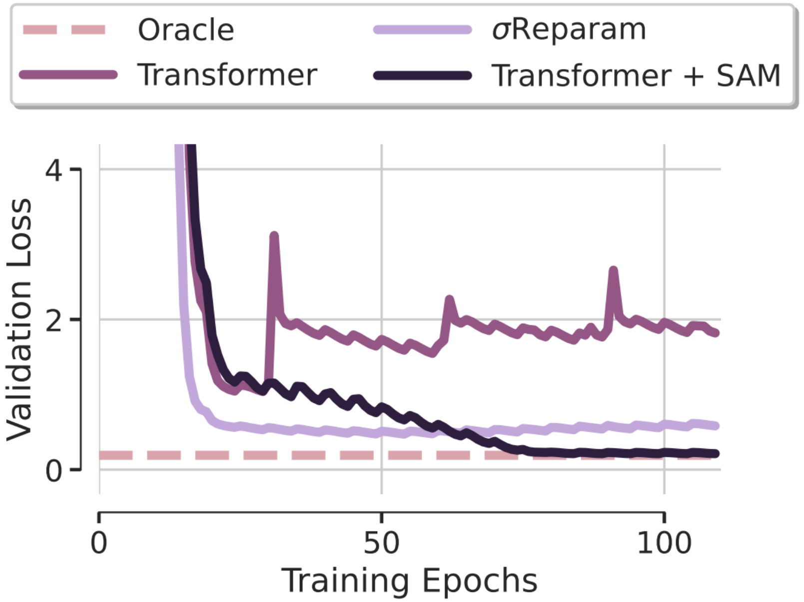
实验4：消融实验——通道注意和时间注意

ETTh1		
	96	$0.381_{\pm0.003}$
	192	$\mathbf{0.409}_{\pm0.002}$
	336	$\mathbf{0.423}_{\pm0.001}$
	720	$\mathbf{0.427}_{\pm0.002}$

ETTm2		
	96	$0.181_{\pm0.005}$
	192	$\mathbf{0.233}_{\pm0.002}$
	336	$\mathbf{0.285}_{\pm0.001}$
	720	$\mathbf{0.375}_{\pm0.001}$

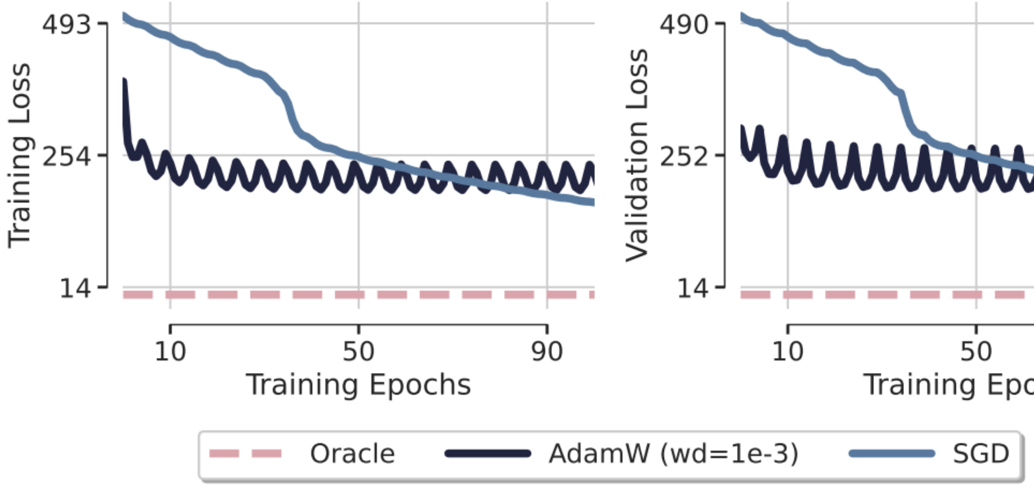
Model	Metrics	H	ETTh1	ETTh2	ETTm1	ETTm2	Electricity	Exchange	Traffic	Weather	**Overall Improvement**
Temporal Attention	MSE	96	$0.496_{\pm0.009}$	$0.401_{\pm0.011}$	$0.542_{\pm0.063}$	$0.330_{\pm0.034}$	$0.291_{\pm0.025}$	$0.684_{\pm0.218}$	$0.933_{\pm0.188}$	$0.225_{\pm0.005}$	12.97%
		192	$0.510_{\pm0.014}$	$0.414_{\pm0.020}$	$0.615_{\pm0.056}$	$0.394_{\pm0.033}$	$0.294_{\pm0.024}$	$0.434_{\pm0.063}$	$0.647_{\pm0.131}$	$0.254_{\pm0.001}$	
		336	$0.549_{\pm0.017}$	$0.396_{\pm0.014}$	$0.620_{\pm0.046}$	$0.436_{\pm0.081}$	$0.290_{\pm0.016}$	$0.473_{\pm0.014}$	$0.656_{\pm0.113}$	$0.292_{\pm0.000}$	
		720	$0.604_{\pm0.017}$	$0.396_{\pm0.010}$	$0.694_{\pm0.055}$	$0.469_{\pm0.005}$	$0.307_{\pm0.014}$	$1.097_{\pm0.084}$	-	$0.346_{\pm0.000}$	
	MAE	96	$0.488_{\pm0.007}$	$0.434_{\pm0.006}$	$0.525_{\pm0.040}$	$0.393_{\pm0.020}$	$0.386_{\pm0.014}$	$0.589_{\pm0.096}$	$0.598_{\pm0.072}$	$0.277_{\pm0.004}$	18.09%
		192	$0.492_{\pm0.010}$	$0.443_{\pm0.015}$	$0.566_{\pm0.032}$	$0.421_{\pm0.019}$	$0.385_{\pm0.014}$	$0.498_{\pm0.033}$	$0.467_{\pm0.072}$	$0.294_{\pm0.001}$	
		336	$0.517_{\pm0.012}$	$0.440_{\pm0.012}$	$0.550_{\pm0.024}$	$0.443_{\pm0.039}$	$0.383_{\pm0.009}$	$0.517_{\pm0.008}$	$0.469_{\pm0.070}$	$0.320_{\pm0.000}$	
		720	$0.556_{\pm0.009}$	$0.442_{\pm0.006}$	$0.584_{\pm0.027}$	$0.459_{\pm0.004}$	$0.396_{\pm0.012}$	$0.782_{\pm0.041}$	-	$0.356_{\pm0.000}$	

06 实验4：消融实验——优化器的选择

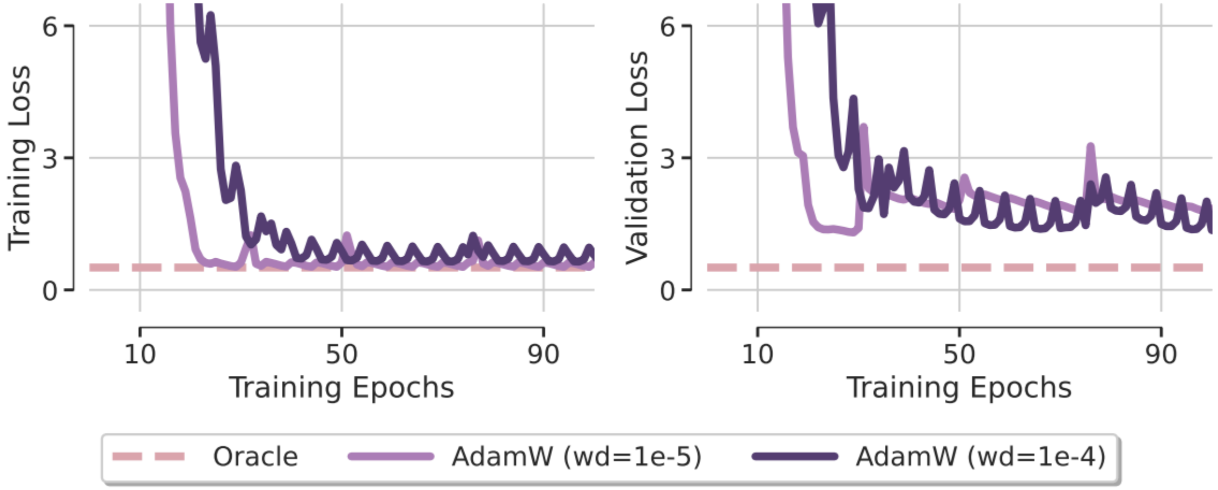


在处理时间序列预测任务时，

<u>通道注意力结合Adam优化器</u>是一个有效且稳健的选择



(a) SGD and AdamW with wd $= 1e{-}3$



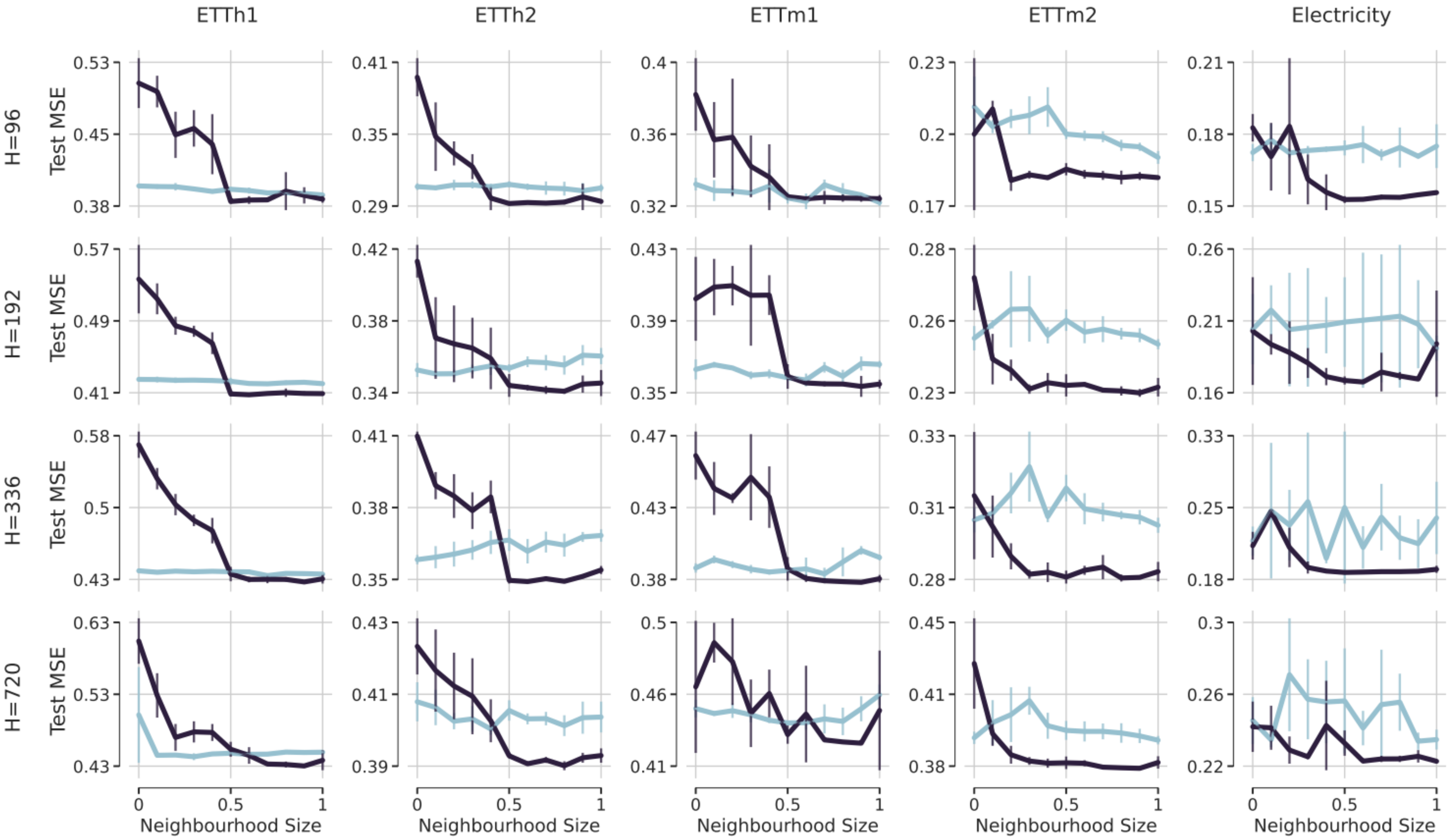
(b) AdamW with wd $\in \{1e{-}5, 1e{-}4\}$.

06

实验5：邻域大小ρ的敏感性

① **平滑行为与稳定性：**当ρ值足够大时，一般在0.7以上，SAMformer能够实现比TSMixer更低的MSE。

② **对比TSMixer的波动性：**TSMixer由于其准线性架构，对ρ的敏感性较低。



谢谢观看

MANY THANKS !

24.6.4