M-pre: 基于 Mamba 的长时序预测模型

苏悦
西安电子科技大学
s3702681@gmail.com
项目链接：https://github.com/Selen-Suyue/M-pre

2025 年 5 月 5 日

摘要

长时序预测（Long-term Time Series Forecasting, LTSF）在能源、金融、交通等多个领域具有重要应用价值，但准确预测长期未来趋势仍然是一个巨大的挑战。现有方法如基于 Transformer 的模型虽然取得了一定成功，但其注意力机制的二次复杂度限制了其处理超长序列的效率。近年来，状态空间模型（State Space Models, SSMs）及其变种 Mamba 因其在线性时间内处理长序列的能力和优异的性能而备受关注。本文提出了一种名为 M-pre 的新型长时序预测模型，该模型核心采用了 Mamba 架构来有效捕捉时间序列中的长程依赖关系。M-pre 通过堆叠 Mamba 块来提取深层时序特征，并结合线性投影层进行最终预测。我们在多个公开的 ETT 数据集上进行了实验，结果表明 M-pre 在多个预测任务上取得了具有竞争力的性能，证明了 Mamba 架构在长时序预测领域的潜力。

关键词: 时序预测, Mamba, 状态空间模型

1 引言

时间序列预测是机器学习中的一个核心问题，旨在根据历史观测数据预测未来的值。其中，长时序预测（LTSF）专注于预测远期未来，这对于许多实际应用至关重要，例如电力负荷规划 [8]、金融市场分析 [11] 和交通流量管理 [10]。

近年来，基于 Transformer 的模型 [8, 11, 9, 4, 7, 10, 6] 在 LTSF 任务中展现了强大的能力。这些模型通过自注意力机制捕捉序列中的依赖关系。然而，标准自注意力机制具有 $O(L^2)$ 的时间和空间复杂度（L 为序列长度），这使得它们在处理极长序列时效率低下且计算成本高昂。为了缓解这个问题，研究者提出了多种稀疏注意力机制或改进的 Transformer 架构 [8, 11]。此外，一些研究也对 Transformer 在时序预测中的有效性提出了质疑，并指出简单的线性模型有时也能达到相当甚至更好的性能 [9, 2]。

与此同时，状态空间模型（SSMs）作为一种经典的序列建模方法重新受到关注。现代 SSMs，特别是结构化 SSMs（S4）及其后续变种 Mamba，通过巧妙的参数化和计算方法（如选择性扫描机制），实现了对长程依赖的高效建模，其计算复杂度与序列长度呈线性关系 $O(L)$。Mamba 在自然语言处理、图像识别等领域取得了显著成功，显示出替代 Transformer 的潜力。

受 Mamba 在线性复杂度下捕捉长程依赖能力的启发，本文探索了将 Mamba 架构应用于长时序预测任务的可行性。我们提出了 M-pre 模型，其核心是堆叠的 Mamba 块，用于从输入时间序列中学习表示。M-pre 旨在结合 Mamba 的效率和建模能力，为 LTSF 提供一个有效的新选择。我们在广泛使用的 ETT（Electricity Transformer Temperature）数据集上评估了 M-pre，并与多种先进的基线模型进行了比较。

本文的主要贡献如下：

- 提出了一种基于 Mamba 架构的新型长时序预测模型 M-pre。
- 在多个 ETT 基准数据集上验证了 M-pre 的有效性，展示了其具有竞争力的预测性能。
- 探索了 Mamba 模型在 LTSF 领域的应用潜力，为该领域的研究提供了新的思路。

2 相关工作

长时序预测领域的研究近年来取得了显著进展，主要可以分为以下几类：

2.1 基于 Transformer 的模型

自 Vaswani 等人提出 Transformer 以来 [14]，其强大的序列建模能力迅速被引入时序预测领域。Autoformer [8] 引入了分解架构和自相关机制来处理趋势性和周期性。FEDformer [11] 在频域利用傅里叶变换和小波变换增强了表示能力。PatchTST [6] 将序列划分为补丁（Patch）并将其作为 Transformer 的输入单元，取得了优异性能。iTransformer [4] 提出了反转 Transformer 结构，将注意力应用于变量维度而非时间维度。Crossformer [10] 设计了跨维度依赖机制来捕捉多变量时序中的变量间关系。TimesNet [7] 将一维时间序列转换为二维张量，利用卷积网络捕捉多周期性。尽管这些模型取得了成功，但如前所述，计算复杂度是它们面临的主要挑战之一。Stationary [5] 等模型则关注于通过序列平稳化来提升 Transformer 的性能。

2.2 基于线性模型的预测

针对 Transformer 复杂性的担忧，一些研究者重新审视了简单线性模型在 LTSF 中的作用。DLinear [9] 提出了一个简单的分解线性模型，在多个基准测试中表现出色。RLinear [2] 进一步研究了线性映射，并提出了残差连接的线性模型。这些工作表明，在某些情况下，简单的线性映射足以捕捉时序的主要动态。

2.3 其他时序模型

除了 Transformer 和线性模型，还有其他类型的深度学习模型被应用于 LTSF。例如，SCINet [3] 使用样本卷积和交互式学习来构建层次化结构。TiDE [1] 提出了一个基于全连接网络的编码器-解码器模型，强调了特征工程的重要性。

2.4 状态空间模型与 Mamba

状态空间模型（SSMs）提供了一种不同的序列建模范式。经典的 SSMs 如卡尔曼滤波器已被广泛使用。近年来，结构化 SSMs（S4）通过特定的参数化（如 HiPPO 理论）和高效算法，使其能够有效处理长序列。Mamba 进一步发展了 SSMs，引入了选择性扫描机制（S6），使得状态转换参数（A, B, C）能够根据输入动态变化，从而实现了更强的建模能力和上下文感知能力。Mamba 在线性时间内完成计算，使其在处理超长序列方面具有显著优势。本文的 M-pre 正是基于 Mamba 架构构建的。

3 方法

3.1 问题定义

长时序预测任务的目标是根据过去 L 个时间步长的观测值 $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \in \mathbb{R}^{L \times N}$，预测未来 T 个时间步长的值 $\mathcal{Y} = \{\mathbf{x}_{L+1}, \ldots, \mathbf{x}_{L+T}\} \in \mathbb{R}^{T \times N}$。其中 N 是变量（特征）的数量。输入 $\mathcal{X}$ 可能还伴随着时间协变量 $\mathcal{M} = \{\mathbf{m}_1, \ldots, \mathbf{m}_{L+T}\} \in \mathbb{R}^{(L+T) \times M}$，其中 M 是协变量特征的数量。模型的任务是学习一个映射函数 f: $(\mathcal{X}, \mathcal{M}_{\text{enc}}, \mathcal{M}_{\text{dec}}) \mapsto \mathcal{Y}$，其中 $\mathcal{M}_{\text{enc}}$ 和 $\mathcal{M}_{\text{dec}}$ 分别是编码器和解码器阶段使用的协变量。

3.2 M-pre 模型结构

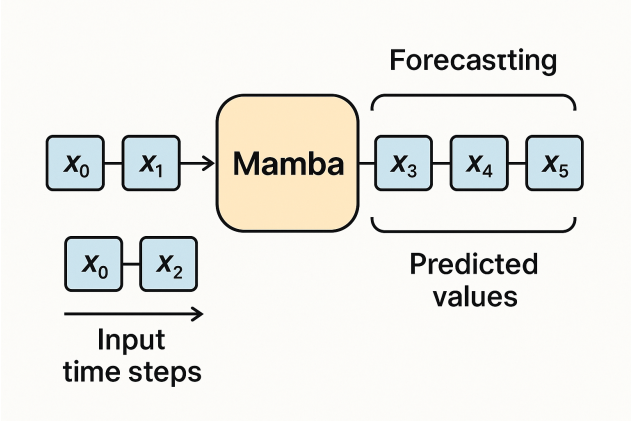

图 1: M-pre 流程示意

我们提出的 M-pre 模型整体结构如图1所示。其核心组件是 Mamba Blocks。

3.2.1 输入处理与归一化

与许多时序模型类似，M-pre 首先对输入序列 $\mathbf{x}_{\text{enc}} \in \mathbb{R}^{B \times L \times N}$ 进行处理（B 为批量大小）。如果配置了使用归一化 ($use_norm = True$)，则采用实例归一化（Instance Normalization）来稳定训练过程：

$$\mu = \frac{1}{L} \sum_{i=1}^{L} \mathbf{x}_{\text{enc},i} \tag{1}$$

$$\sigma^2 = \frac{1}{L} \sum_{i=1}^{L} (\mathbf{x}_{\text{enc},i} - \mu)^2 \tag{2}$$

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x}_{\text{enc}} - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{3}$$

其中 $\mu, \sigma \in \mathbb{R}^{B \times 1 \times N}$ 是每个实例每个变量的均值和标准差。归一化后的序列 $\mathbf{x}_{\text{norm}}$ 将用于后续处理。预测结果在输出前会进行反归一化。模型还将时间协变

量 $\mathbf{x}_{\text{mark\_enc}} \in \mathbb{R}^{B\times L\times M}$ 与 $\mathbf{x}_{\text{enc}}$（或 $\mathbf{x}_{\text{norm}}$）在特征维度上拼接，形成 Mamba 模块的输入。输入数据随后被调整维度顺序以匹配 Mamba 块的期望输入格式 $((batch, seq\_len, feature))$。

3.2.2 Mamba 核心模块

M-pre 的核心是堆叠的 Mamba 层。模型 $Mamba$ 包含三个 $MambaBlock$。每个 $MambaBlock$ 包含以下组件：

1. **输入归一化 (RMSNorm):** 对输入进行 RMSNorm，这是一种简化的层归一化。
2. **输入投影:** 一个线性层将输入维度 d_{model} 扩展到 $2 \times d_{\text{model}}$。
3. **一维卷积:** 应用一个深度可分离的一维卷积（$nn.Conv1d$，kernel_size=3, padding=1），用于捕捉局部信息。
4. **SiLU 激活:** 应用 SiLU（Sigmoid Linear Unit）激活函数。
5. **S6 模块 (选择性状态空间模型):** 这是 Mamba 的核心。它接收卷积层的输出，并通过选择性扫描机制进行序列建模。S6 模块内部通过线性层动态计算状态空间参数 $\mathbf{A}, \mathbf{B}, \mathbf{C}$ 以及离散化步长 Δ。状态更新过程可以概括为：

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}) \tag{4}$$

$$\bar{\mathbf{B}} = \Delta\mathbf{B} \tag{5}$$

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t \tag{6}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t \tag{7}$$

   其中 $\mathbf{x}_t$ 是当前输入，$\mathbf{h}_t$ 是隐藏状态，$\mathbf{y}_t$ 是输出。Mamba 的关键在于 $\Delta, \mathbf{B}, \mathbf{C}$ 是根据输入 $\mathbf{x}_t$ 动态生成的，使得模型具有选择性。代码中的 $S6$ 类实现了这一过程，包括参数计算和离散化。
6. **SiLU 激活:** 再次应用 SiLU 激活函数。
7. **残差连接:** 原始输入 $\mathbf{x}$ 经过一个线性层 $\mathbf{D}$ 和 SiLU 激活函数，形成残差分支。
8. **门控机制:** S6 模块的输出与残差分支的输出逐元素相乘。
9. **输出投影:** 另一个线性层将维度从 $2 \times d_{\text{model}}$ 映射回 d_{model}。

通过堆叠多个这样的 $MambaBlock$，模型可以学习到更复杂的时序依赖关系。

3.2.3 输出层

经过多个 Mamba 块处理后，得到的表示 $\mathbf{h}_{\text{out}} \in \mathbb{R}^{B\times L'\times d_{\text{model}}}$ 被送入一个最终的线性投影层 $projector$，该层将 Mamba 的输出映射到所需的预测长度 T：

$$\mathbf{y}_{\text{pred}} = \text{Projector}(\mathbf{h}_{\text{out}}) \in \mathbb{R}^{B\times T\times N} \tag{8}$$

最终，模型输出 $out[:, -self.pred_len:, :]$，确保输出长度为 T。

4 实验

4.1 数据集

我们在四个广泛使用的 ETT（Electricity Transformer Temperature）数据集上评估了 M-pre 的性能：ETTm1, ETTm2, ETTh1, ETTh2。这些数据集记录了电力变压器的油温和其他相关指标，具有明显的多尺度周期性和趋势性，是评估 LTSF 模型能力的常用基准。我们遵循标准设置，使用均方误差（MSE）和平均绝对误差（MAE）作为评估指标。

4.2 基线模型

我们将 M-pre 与一系列先进的 LTSF 模型进行了比较，包括：

- **基于 Transformer 的模型:** iTransformer [4], PatchTST [6], Crossformer [10], TimesNet [7], FEDformer [11], Autoformer [8], Stationary [5]。
- **基于线性模型的模型:** RLinear [2], DLinear [9]。
- **其他模型:** TiDE [1], SCINet [3], U-preV1, U-preV2 (https://github.com/Selen-Suyue/U-pre)。

4.3 实现细节

我们基于提供的 Python 代码实现了 M-pre 模型。模型使用 PyTorch 框架构建，并在 NVIDIA GPU 上进行训练和测试。我们使用 Adam 优化器进行模型训练。具体的超参数，如学习率、批量大小（batch_size）、状态维度（state_size）、模型维度（d_model）等，依据代码中的配置或常见的经验值设定。我们对所有数据集以 96 的预测长度进行了实验，并报告了平均结果。模型的输入序列长度 L 和预测序列长度 T (pred_len) 根据标准基准设置。

模型	ETTm1 (MSE/MAE)	ETTm2 (MSE/MAE)	ETTh1 (MSE/MAE)	ETTh2 (MSE/MAE)	平均 (MSE/MAE)
iTransformer	0.334/0.368	0.180/0.264	0.386/0.405	0.297/0.349	0.299/0.347
RLinear	0.355/0.376	0.182/0.265	0.386/0.395	0.288/0.338	0.303/0.344
PatchTST	0.329/0.367	0.175/0.259	0.414/0.419	0.302/0.348	0.305/0.348
Crossformer	0.404/0.426	0.287/0.366	0.423/0.448	0.340/0.374	0.364/0.404
TiDE	0.364/0.387	0.207/0.305	0.384/0.402	0.340/0.374	0.324/0.367
TimesNet	0.338/0.375	0.187/0.267	0.479/0.464	0.400/0.440	0.351/0.387
DLinear	0.345/0.372	0.193/0.292	0.386/0.400	0.402/0.414	0.332/0.369
SCINet	0.418/0.438	0.286/0.377	0.654/0.599	0.376/0.419	0.434/0.458
FEDformer	0.379/0.419	0.203/0.287	0.513/0.491	0.449/0.459	0.386/0.414
Stationary	0.386/0.398	0.192/0.274	0.449/0.459	0.526/0.516	0.388/0.412
Autoformer	0.505/0.475	0.255/0.339	0.449/0.459	0.450/0.459	0.415/0.433
U-preV1	0.466/0.451	0.195/0.275	0.524/0.483	0.367/0.395	0.388/0.401
U-preV2	0.370/0.396	0.188/0.273	0.419/0.429	0.325/0.371	0.326/0.367
M-pre (本文)	**0.334/0.377**	**0.186/0.268**	**0.412/0.427**	**0.366/0.397**	**0.325/0.367**

表 1: ETT 数据集上的长时序预测结果 (MSE/MAE)。数值越低越好。

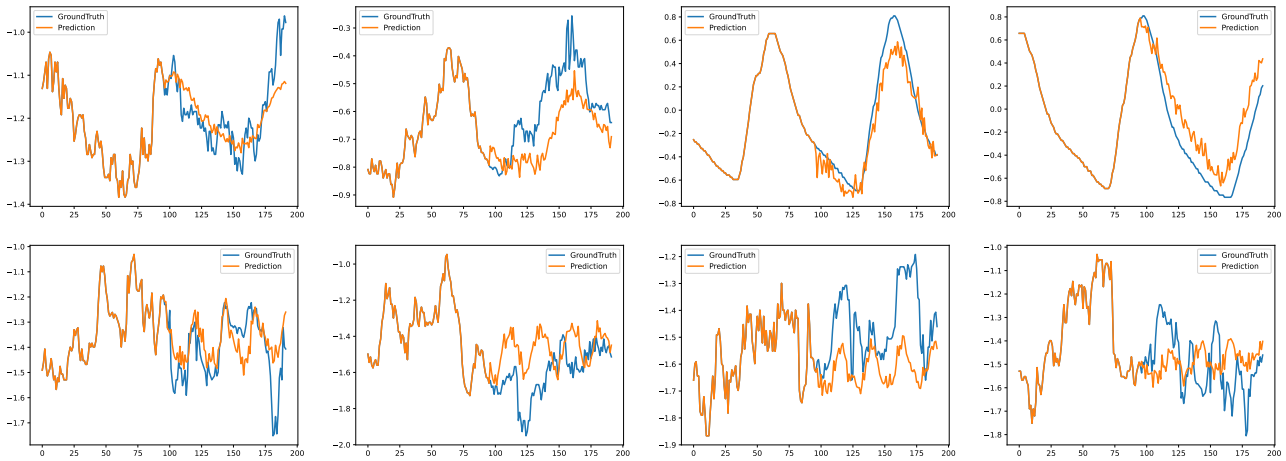

图 2: M-pre 模型下的预测结果可视化示例

4.4 实验结果

我们在 ETT 数据集上的长时序预测的主要结果如表 1 所示。表中数值为 MSE/MAE，越低越好。部分预测示例如图 2所示。

4.5 结果分析

从表 1 可以看出，M-pre 模型在多个 ETT 数据集上取得了具有竞争力的结果。

- 在 ETTm1 数据集上，M-pre 的 MSE（0.334）与表现最好的 iTransformer 持平，MAE（0.377）略高于 iTransformer 和 PatchTST。
- 在 ETTm2 数据集上，M-pre 的 MSE（0.186）和 MAE（0.268）均表现良好，仅次于 PatchTST 和 iTransformer 等少数模型。
- 在 ETTh1 和 ETTh2 数据集上，M-pre 的表现相对稳定，优于部分复杂的 Transformer 模型（如 FEDformer, Autoformer, SCINet），但略逊于 iTransformer 和 RLinear 等顶级模型。
- 在所有四个数据集的平均性能上，M-pre 的 MSE（0.325）和 MAE（0.367）与 TiDE、U-preV2 等模型

处于同一水平，优于许多经典的 Transformer 模型，但略低于 iTransformer、RLinear 和 PatchTST。

总体而言，实验结果初步验证了 Mamba 架构在长时序预测任务中的潜力。M-pre 作为一个基于 Mamba 的直接实现，能够在不进行过多针对性优化的情况下，达到与许多先进模型相当甚至更好的性能，特别是在计算效率上具有潜在优势（线性复杂度）。这表明 Mamba 的选择性状态空间机制能够有效捕捉时序数据中的复杂动态。

5 结论

本文提出了一种新的长时序预测模型 M-pre，该模型利用了最新的 Mamba 架构（一种选择性状态空间模型）来捕捉时间序列中的长程依赖关系。Mamba 的核心优势在于其线性计算复杂度以及通过选择性机制动态调整参数的能力。我们在标准的 ETT 基准数据集上对 M-pre 进行了评估，并与多种流行的基线模型进行了比较。实验结果表明，M-pre 在多个数据集上取得了具有竞争力的预测精度，证明了 Mamba 架构应用于长时序预测任务的可行性和潜力。

未来的工作可以从以下几个方面展开：

- 对 M-pre 的模型结构进行更深入的优化，例如探索不同数量的 Mamba 块、不同的状态空间维度、以及更适合时序任务的输入/输出处理方式。
- 将 Mamba 与其他时序建模技术（如序列分解、频域分析）相结合。
- 在更广泛的数据集和应用场景中验证 M-pre 的性能和泛化能力。
- 进一步分析 Mamba 的选择性机制如何在时序数据中发挥作用。

我们相信，基于 Mamba 等高效序列模型的时序预测方法将在未来得到更多关注和发展。

参考文献

[1] Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan K Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tiDE: Time-series dense encoder. *Transactions on Machine Learning Research*, 2023.

[2] Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping, 2023. arXiv preprint arXiv:2305.10721.

[3] Minhao LIU, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia LAI, Lingna Ma, and Qiang Xu. SCINet: Time series modeling and forecasting with sample convolution and interaction. In *Advances in Neural Information Processing Systems*, 2022.

[4] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting, 2024. *International Conference on Learning Representations (ICLR)*.

[5] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, 2022.

[6] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[7] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[8] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with Auto-Correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

[9] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2023.

[10] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency

for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[11] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proc. 39th International Conference on Machine Learning (ICML)*, 2022.

[12] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

[13] Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752, 2023.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017.