一种用于自动和弦识别的基于 Mamba 的模型

Chunyu Yuan纽约市
立大学研究生中心计算机
科学

Johanna Devaney布鲁克林学院和纽约市
立大学研究生中心数据分析与可视化及音乐
johanna.devaney@brooklyn.cuny.edu

arXiv:2601.02101v1 [cs.SD] 5 Jan 2026

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

摘要

在这项工作中，我们提出了一种高效的新方案，即名为 BMACE（用于自动和弦估计的双向 Mamba 网络）的基于 Mamba 的模型。该模型在双向 Mamba 层中采用选择性结构化状态空间模型，以有效建模时间依赖关系。我们的模型取得了与最先进模型相当的高预测性能，同时所需参数更少，计算资源也更低。

1. 引言与背景

自动和弦识别/估计（ACE）在音乐信息检索（MIR）研究中已有悠久历史 [1]。尽管现代深度学习技术的应用带来了显著改进 [2], , 但即使是近期最先进的方法，性能仍然存在上限 [3–5]。此前已指出，自动和弦估计面临的一些挑战包括大量的标签排列 [3], 以及专家标注人员之间的意见分歧 [6–9], , 而罕见和弦的情况尤其如此 [8,10]。基于 Transformer 的模型（例如， [4] ）擅长捕捉自动和弦估计任务所需的时间依赖关系，但由于其复杂度相对于输入长度呈二次增长，也会带来显著的计算开销。Transformer 架构日益增加的复杂度，加上其对内存和处理能力的高要求，限制了它们在低延迟环境中的适用性，例如实时音乐分析系统或嵌入式设备。这类应用要求在模型准确率与计算效率之间仔细权衡。

本文评估了通过更新模型架构，具体而言，在双向 Mamba 层中加入选择性结构化状态空间模型，能够为自动和弦估计任务带来多大的性能提升和模型精简。具体来说，我们提出了 BMACE（用于自动和弦估计的双向 Mamba 网络），这是一种新颖的基于 Mamba 的模型，在双向 Mamba 层中融入了选择性结构化状态空间模型，以增强对时间依赖关系的建模。值得注意的是，该模型以更少的参数和更低的计算成本，取得了与前代模型相当的性能。

© C. Yuan 和 J. Devaney。依据知识共享署名 4.0 国际许可协议（CC BY 4.0）授权。**署名：** C. Yuan 和 J. Devaney，《基于 Mamba 的自动和弦识别模型》，载于第 *25* 届国际音乐信息检索学会会议迟发演示环节扩展摘要，美国旧金山，2024 年。

2. 双向 Mamba 网络

受双向 Transformer 启发，我们提出了一种专为和弦估计/识别设计的轻量级双向 Mamba 网络：BMACE（基于双向 Mamba 的自动和弦估计网络）。Mamba 架构于 2023 [11] 年末首次提出，自发布以来迅速受到关注。它已应用于一些语音[12–14] 和音乐信息检索 [15,16] 任务，但尚未用于自动和弦估计。Mamba 有别于其他模型之处在于，它摒弃了常见的注意力机制和多层感知机模块，采用更精简的方案。因此，该模型不仅更轻、更快，还具备随序列长度线性扩展的独特能力，这一特点使其超越了前代模型。Mamba 设计的核心是其选择性状态空间（SSM）：这类循环模型根据当前输入选择性地处理信息，从而有效滤除无关数据，聚焦于对高效处理最关键的内容。此外，Mamba 以单一、统一的状态空间模型模块取代 Transformer 中复杂的注意力机制和多层感知机模块，简化了架构，提升了推理速度并降低了计算负载。Mamba 还采用硬件感知并行，通过专门设计的并行算法优化循环运算，提高硬件效率，进而有望进一步提升性能。

图1展示了我们所实验的基于 Mamba 的模型的三种变体。第一种（MACE-V）是包含两个纵向 Mamba 层的原始 Mamba 模型。第二种（MACE-H）是由两个串联的／横向模型构成的 Mamba 模型。第三种是我们提出的模型 BMACE，它展示了双向 Mamba 网络的结构。双向 Mamba 块和全连接层是该网络的主要模块。模型将 10 秒音频信号处理为常数 Q 变换（CQT）特征。模型在输入端整合了一个全连接层，随后将输入送入两个掩码方向相反的 Mamba 块，在图1中以虚线框表示。这些模块的输出会拼接起来，再经过一个全连接层，以保持输入的原始维度。

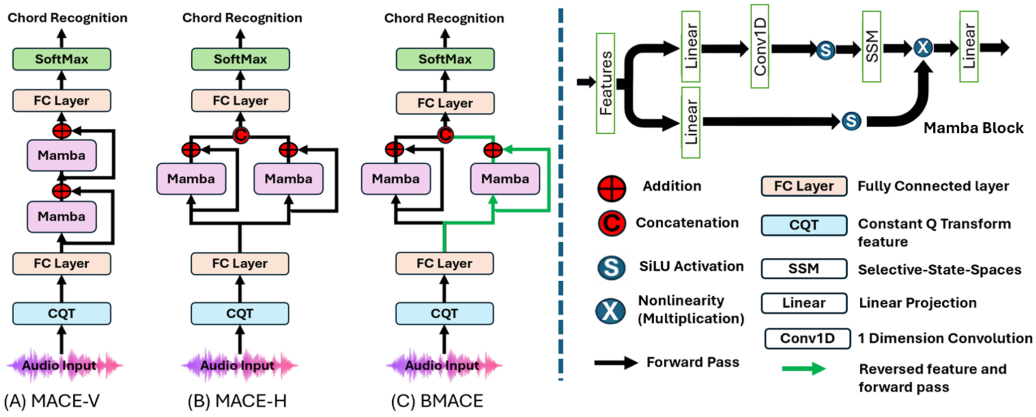


图1。采用 Mamba块的和弦识别模型架构。该图展示了三种变体：MACE-V（A）、MACE-H（B）和 BMACE（C）。每种模型都利用 Mamba块改进特征处理。Mamba块采用选择性状态空间（SSM）、SiLU激活和一维卷积（Conv1D）进行特征变换。图中突出显示了各模型中的前向传播操作、特征反转、加法和拼接机制，并通过全连接（FC）层生成 SoftMax输出，用于和弦识别。

模型	大调-小调标签类型				大词汇表标签类型								
	根音↑	大调-小调↑	GFlops↓	参数量↓	根音↑	三度音↑	三和弦↑	七和弦↑	四音和弦↑	大调-小调↑	MIREX↑	GFlops↓	参数量↓
CRNN [3]	0.8185	**0.7796**	0.0957	435,609	0.8026	0.7459	0.6384	0.6426	0.5448	**0.7544**	0.7543	0.1038	472,874
BTC [4]	0.8202	0.7628	0.6282	2,910,361	**0.8051**	**0.7524**	**0.6469**	0.6506	**0.5604**	0.7531	0.7486	0.6362	2,929,066
MACE-V	0.7920	0.7309	**0.0247**	**111,161**	0.7739	0.715	0.6084	0.6137	0.5242	0.7166	0.7057	**0.0328**	**129,866**
MACE-H	0.7898	0.7347	0.0261	114,361	0.7833	0.7211	0.6188	0.6236	0.5314	0.7304	0.7238	0.0422	151,626
BMACE	**0.8212**	0.7678	0.0261	114,361	0.8043	0.7455	0.6426	**0.6526**	0.5571	0.7536	**0.7595**	0.0422	151,626

表 1。CRNN [3], BTC [4], 和我们的三种 Mamba 变体（MACE-V、MACE-H 和 BMACE）在 uspop2002 数据集上的表现，其加权和弦符号召回率（WCSR）得分。

我们在这些块和层中添加了残差操作，以增加信息熵。

3. 实验

3.1 实验设置

我们的模型使用 Pytorch [17] 框架实现。所有实验均在 Lambda [1] 实例节点上进行，该节点配备一块 NVIDIA RTX A6000 GPU（24 GB）、14 个虚拟 CPU、46 GiB 内存和 512 GiB 固态硬盘。我们的模型使用 uspop2002 数据集 [18]的 MARL 标注 [2] 进行训练和验证。每个 10 秒音频信号都经过处理，相邻信号之间有 5 秒的重叠。信号以 22,050 Hz 采样，并使用常数 Q 变换（CQT）进行分析；该变换涵盖从 C1 开始的 6 个八度，每八度频点数为 24，步长为 2048。随后，使用公式 $S_{\log} = \ln(S+\epsilon)$ 将 CQT 特征转换为对数振幅，其中 S 表示 CQT 特征，而 ϵ 是一个极小的数。接下来应用全局 Z 标准化，使用根据训练数据计算得出的均值和方差。

我们将第 2 节所述模型的三个版本（MACE-V、MACE-H 和 BMACE），与当前最先进的基于 CRNN [3] 和基于 Transformer [4]的模型进行比较，评估其在 25 标签大调-小调和 170 标签大型和弦词汇表上的表现。

3.2 结果

表 1 展示了模型验证结果。BMACE 在某些标签类型上的表现略优于 CRNN 和 BTC 模型，不过这种差异可能并不具有统计显著性。然而，与非双向 Mamba 模型（MACE-V 和 MACE-H）相比，性能有明显提升。如表 1 所示，各模型在规模和处理需求方面的差异最为显著。正如先前观察到的，CRNN 比基于 Transformer 的模型更高效，而 Mamba 网络同样更高效这一说法也成立。所有基于 Mamba 的模型所用参数仅为基于 Transformer 的 BTC 模型的 1/25，且比 CRNN 小略少于 1/3。参数数量的减少也体现在运行模型所需的 GFlops 更低。

[1] https://cloud.lambdalabs.com/instances[2] https://github.com/tmc323/Chord-Annotations

4. 致谢

本研究得到美国国家科学基金会（NSF）资助，项目编号为 2228910。本研究中表达的观点仅代表作者个人意见，并不一定反映 NSF 的观点。

5. 参考文献

[1] J. Pauwels、K. O'Hanlon、E. Gómez 和 M. B. San- dler，“音频自动和弦识别二十年”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，荷兰代尔夫特，2019年，第54–63页。[2] F. Korzeniowski 和 G. Widmer，“用于音乐和弦识别的全卷积深度听觉模型”，载于*2016*年电气电子工程师学会第*26*届信号处理机器学习国际研讨会（*MLSP*）。电气电子工程师学会，2016年，第1–6页。[3] B. McFee 和 J. P. Bello，“面向大词汇量和弦识别的结构化训练”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，中国苏州，2017年，第188–194页。[4]J. Park、K. Choi、S. Jeon、D. Kim 和 J. Park，“用于音乐和弦识别的双向变换器”，载于国际音乐信息检索学会会议论文集（*ISMIR*），A. Flexer、G. Peeters、J. Urbano 和 A. Volk 编，2019年，第620–627页。 [在线]。可访问：
h
t
t
p://
archives.ismir.net/ismir2019/paper/000075.pdf
[5] T. Ito 和 S. Arai，“用于CNN-LSTM自动和弦识别的谐波表示”，载于国际控制论与智能系统会议（*ICORIS*）。电气电子工程师学会，2021年，第1–5页。[6] E. J. Humphrey 和 J. P. Bello，“关于自动和弦估计的四点及时见解”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，西班牙马拉加，2015年，第673–679页。[7] N. Condit-Schultz、Y. Ju 和 I. Fujinaga，“自动谐波分析的灵活方法：巴赫和普雷托里乌斯众赞歌的多重标注”，载于国际音乐信息检索学会会议（*ISMIR*）论文集，2018年，第66–73页。[8]H. V. Koops、W. B. de Haas、J. A. Burgoyne、J. Bransen、A. Kent-Muller 和 A. Volk，“流行音乐和声标注中的标注者主观性”，《新音乐研究杂志》，第48卷，第3期，第232–252页，2019年。[9]H. Koops、W. de Haas、J. Bransen 和 A. Volk，“通过深度学习集成谐波音程表示实现和弦标签个性化”，载于
第一届深度学习与音乐国际会议论文集，2017年，第19–25页。[10] Y. Ni、M. McVicar、R. Santos-Rodríguez 和 T. D. Bie，“理解主观性对和弦估计准确率测量的影响”，*IEEE*音频、语音与语言处理汇刊，第21卷，第12期，第2607–2615页，2013年。[11] A. Gu 和 T. Dao，“Mamba：利用选择性状态空间进行线性时间序列建模”，*arXiv*预印本 *arXiv:2312.00752*，2023年。[12]X. Jiang、C. Han 和 N. Mesgarani，“ 双路径 Mamba：用于语音分离的短期和长期双向选择性结构化状态空间模型”，*arXiv*预印本 *arXiv:2403.18257*，2024。[13] K. Li 和 G. Chen，“Spmamba：语音分离只需状态空间模型”，*arXiv*预印本 *arXiv:2404.02063*，2024。[14] C. Quan 和 X. Li，“面向静止和移动说话人的多通道长期流式神经语音增强”，*arXiv*预印本 *arXiv:2403.07675*，2024。[15] J. Bai、Y. Fang、J. Wang 和 X. Zhang，“用于音乐分离的两阶段频带拆分 Mamba-2网络”，*arXiv*预印本 *arXiv:2409.06245*，2024。[16] J. Chen、T. Xie、X. Tang、J. Wang、W. Dong 和 B. Shi，“Musicmamba：一种以调式精准度生成中国传统音乐的双特征建模方法”，*arXiv*预印本 *arXiv:2409.02421*，2024。[17] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga等，“PyTorch：一种命令式风格的高性能深度学习库”，神经信息处理系统进展，第32卷，2019年。[18] A. Berenzweig、B. Logan、D. P. Ellis 和 B. Whitman，“对声学与主观音乐相似度度量的大规模评估”，《计算机音乐杂志》，第63–76页，2004年。