软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):1887−1902 [doi: 10.13328/j.cnki.jos.007542] [CSTR: 32375.14.jos.007542]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合音乐知识结构化表征的高精度符号音乐理解*

黄恒焱[1], 邹 逸[2], 时乐轩[3], 程皓楠[2], 叶 龙[1,4]

[1](中国传媒大学 数据科学与智能媒体学院, 北京 100024)
[2](媒体融合与传播国家重点实验室 (中国传媒大学), 北京 100024)
[3](中国传媒大学 音乐与录音艺术学院, 北京 100024)
[4](媒介音视频教育部重点实验室 (中国传媒大学), 北京 100024)
通信作者: 程皓楠, E-mail: haonancheng@cuc.edu.cn

摘 要: 符号音乐理解 (symbolic music understanding, SMU) 是多媒体内容理解的重要任务之一, 旨在从符号化音乐表示中提取旋律、力度、作曲家风格、情感与流派等多维音乐属性. 现有方法在音乐序列依赖建模方面取得了显著进展, 但是仍然存在两方面关键问题: (1) 表示单一化: 将复杂的音乐结构简化为线性符号序列, 忽略了音乐固有的多维层级信息; (2) 乐理知识缺乏: 基于序列数据驱动的模型难以融入系统化乐理知识, 限制了对音乐深层语义的理解. 针对上述问题, 提出了一种融合音乐知识结构化表征的高精度符号音乐理解模型 CNN-Midiformer. 该模型首先基于音乐理论构建音乐知识和音乐序列的结构化表征; 其次, 设计互补音乐特征提取模块, 利用卷积神经网络 (convolutional neural network, CNN) 提取音乐知识结构化表征的深层局部特征, 并通过 Transformer 编码器的自注意力机制捕获音乐序列的深层语义特征; 最后, 设计音乐知识自适应增强的特征融合模块, 利用高效的交叉注意力机制将 CNN 提取的深层音乐知识特征与 Transformer 编码器的深层语义特征进行动态融合, 实现对序列语境的感知与特征增强. 在 6 个公开符号音乐理解数据集 Pop1K7、ASAP、POP909、Pianist8、EMOPIA 和 ADL 上的对比实验表明, 所提出的模型 CNN-Midiformer 在旋律识别、力度预测、作曲家分类、情感分类和流派分类这 5 个符号音乐理解的基准下游任务上均优于最新方法, 相较于基线模型准确率提升 0.21–7.14 个百分点.

关键词: 符号音乐理解; 音乐知识结构化表征; 卷积神经网络; Transformer; 特征融合
中图法分类号: TP37

中文引用格式: 黄恒焱, 邹逸, 时乐轩, 程皓楠, 叶龙. 融合音乐知识结构化表征的高精度符号音乐理解. 软件学报, 2026, 37(5): 1887–1902. http://www.jos.org.cn/1000-9825/7542.htm

英文引用格式: Huang HY, Zou Y, Shi LX, Cheng HN, Ye L. High-precision Symbolic Music Understanding Incorporating Structured Representation of Music Knowledge. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 1887–1902 (in Chinese). http://www.jos.org.cn/1000-9825/7542.htm

High-precision Symbolic Music Understanding Incorporating Structured Representation of Music Knowledge

HUANG Heng-Yan[1], ZOU Yi[2], SHI Le-Xuan[3], CHENG Hao-Nan[2], YE Long[1,4]

[1](School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China)
[2](State Key Laboratory of Media Convergence and Communication (Communication University of China), Beijing 100024, China)
[3](School of Music and Recording Arts, Communication University of China, Beijing 100024, China)
[4](Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, Beijing 100024, China)

* 基金项目: 国家自然科学基金 (62201524)
本文由“多媒体智能理解与生成”专题特约编辑孙立峰教授、闵巍庆副研究员、马占宇教授、蒋树强研究员、彭宇新教授、田丰研究员、黄庆明教授推荐.
收稿时间: 2025-05-26; 修改时间: 2025-07-11, 2025-08-30; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23
CNKI 网络首发时间: 2025-12-26

Abstract: Symbolic music understanding (SMU) is a crucial task in multimedia content understanding, aiming to extract multi-dimensional musical attributes such as melody, dynamics, compositional style, emotion, and genre from symbolic representations. Although existing approaches have substantially advanced dependency modeling in musical sequences, two critical challenges remain: (1) Simplified representation: Current methods typically flatten complex musical structures into linear symbolic sequences, overlooking the inherent multi-dimensional hierarchical information; (2) Lack of music-theory integration: Purely data-driven sequence models struggle to incorporate structured music-theory knowledge, limiting deep semantic understanding of music. To address these issues, this study proposes CNN-Midiformer, a high-precision symbolic music understanding model that integrates structured representations of musical knowledge. First, the model constructs structured representations for music theory and musical sequences based on domain knowledge. Second, a complementary music-feature extraction module is devised to employ convolutional neural networks (CNN) for capturing deep local features from structured musical-knowledge representations, while a Transformer encoder with self-attention captures deep semantic features from musical sequences. Finally, a music-knowledge adaptive-enhancement feature-fusion module dynamically integrates the deep musical-knowledge features extracted by CNN with the deep semantic features of the Transformer via an efficient cross-attention mechanism, thus enhancing contextual sequence understanding and representation learning. Comparative experiments conducted on six public symbolic-music datasets (Pop1K7, ASAP, POP909, Pianist8, EMOPIA, and ADL) demonstrate that CNN-Midiformer surpasses state-of-the-art methods across five benchmark downstream tasks: melody recognition, dynamics prediction, composer classification, emotion classification, and genre classification, achieving a precision gain of 0.21–7.14 percentage points over baseline models.

Key words: symbolic music understanding (SMU); structured representation of music knowledge; convolutional neural network (CNN); Transformer; feature fusion

符号音乐理解 (symbolic music understanding, SMU) 是多媒体内容理解的重要任务之一, 不同于直接处理 wav、mp3 等音频格式的音乐理解任务, SMU 旨在从乐器数字接口 (musical instrument digital interface, MIDI)、MusicXML 等符号化音乐表示中提取旋律[1,2]、力度[3,4]、作曲家风格[5,6]与情感[7,8]等多维音乐属性, 如图 1 所示. 符号化音乐表示一方面能够精确记录音高、时值、力度、音轨和乐器等构成音乐骨架的基础要素, 为深层计算分析提供了结构化数据基础[9]; 另一方面避免了在处理音频时, 自动音乐转录 (automatic music transcription, AMT) 可能引入的识别误差, 使模型能更专注于音乐内容本身的深层理解. 因此, SMU 在音乐自动作曲与编曲、旋律补全、智能乐谱分析等领域有着广泛的应用.

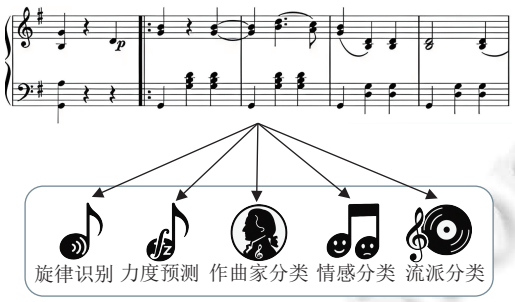


图 1 符号音乐理解任务概览图

基于 Transformer 架构及其变体 (BERT[10]、GPT[11]等) 的 SMU 方法通过大规模预训练模型和自注意力机制, 能有效地捕捉符号音乐序列中的长距离依赖关系, 在旋律识别、力度预测、作曲家分类与情感分类等任务上实现了显著性能提升, 极大地推动了符号音乐理解领域的发展[12,13]. 然而, 音乐并非简单的符号序列, 还具有复杂的多维层级信息[14]. 基于 Transformer 架构的模型在处理序列数据方面表现出色, 但仅以线性序列视角处理音乐, 忽略了音乐的结构和层级, 导致模型难以有效建模音乐和声、旋律和节奏等基本元素间的关系, 限制了对音乐内在规律与模式的理解. 因此, 当前方法仍主要存在两方面问题: (1) 表示单一化: 将复杂的音乐结构简化为线性符号序列, 忽略了音乐固有的多维层级信息; (2) 乐理知识缺乏: 基于序列数据驱动的模型难以融入系统化乐理知识, 限制了对音乐深层语义的理解.

音乐学研究表明, 许多关键的音乐结构特征可以通过数学建模进行有效量化和表征[15]. 例如, 音高类别转移

矩阵 (pitch class transition matrix, PCTM) 能够统计性地刻画旋律在不同音高间的转移偏好, 音符长度转移矩阵 (note length transition matrix, NLTM) 能捕捉节奏的转换规律, 结合两种转移矩阵可以反映不同流派的风格[16]. 这些结构化特征兼具音乐学意义与统计特性, 蕴含了关于反映旋律和节奏的重要局部结构信息, 在序列建模的协同下能够为模型提供乐理指导, 有助于提升模型在下游任务上的精度.

基于该思路, 本文提出了一种融合音乐知识和音乐序列的深层符号音乐理解模型 CNN-Midiformer, 实现了高精度的符号音乐理解. 不同于完全基于 Transformer 架构的符号音乐理解模型, 本文首先提出了互补音乐特征提取模块, 利用卷积神经网络 (convolutional neural network, CNN) 从音乐知识结构化表征中提取深层音乐知识特征, 并基于 Transformer 架构捕捉符号音乐序列的长距离序列依赖关系; 其次, 设计了音乐知识自适应增强的特征融合模块, 利用高效的交叉注意力机制将 CNN 提取的深层音乐知识特征与 Transformer 编码器的深层语义特征进行自适应融合, 实现了对序列语境的感知与特征增强. 本文的主要贡献可概括如下.

- 提出了一种深层符号音乐理解模型 CNN-Midiformer, 通过 CNN 与 Transformer 编码器协同学习音乐知识和音乐序列, 实现了高精度的符号音乐理解.
- 设计了互补音乐特征提取模块, 利用 CNN 提取音乐知识特征中蕴含的音乐局部结构信息, 将音乐知识与 Transformer 编码器的深层语义特征融合, 增强了模型对音乐序列的建模能力.
- 设计了音乐知识自适应增强的特征融合模块, 通过高效的交叉注意力机制将 CNN 提取的深层音乐知识特征根据序列语境动态地增强 Transformer 编码器的表示, 实现了多源音乐特征的自适应融合.
- 在 6 个公开符号音乐理解数据集 Pop1K7、ASAP、POP909、Pianist8、EMOPIA 和 ADL 上的对比实验表明, 该模型在旋律识别、力度预测、作曲家分类、情感分类和流派分类这 5 个符号音乐理解的基准下游任务上均优于最新方法, 相较于基线模型准确率提升 0.21–7.14 个百分点.

本文第 1 节回顾符号音乐理解及相关深度学习模型的研究现状. 第 2 节详细阐述 CNN-Midiformer 模型的架构设计, 包括模型总框架、互补音乐特征提取模块和音乐知识自适应增强的特征融合模块. 第 3 节介绍实验设置, 并展现详细的实验结果与分析. 最后对全文进行总结.

1 符号音乐理解的相关工作

符号音乐理解 (SMU) 融合了音乐学、计算机科学及自然语言处理 (natural language processing, NLP) 等多学科的理论与技术[17]. 然而, 音乐不仅具有音高、时值、力度等多维同步特征, 还呈现从音符到乐句、乐段、乐章的层级结构, 与自然语言序列存在本质差异[14]. 直接基于现有 NLP 方法无法充分捕捉音乐数据的独特特征, 这要求研究者面向音乐属性构建专门的模型与算法.

早期研究基于 Word2Vec[18,19]等词嵌入方法, 将音符视为“词汇”并学习其分布式表示, 进而捕捉音符间的语义关系, 在旋律识别等任务中取得初步成果. 隐马尔可夫模型 (hidden Markov model, HMM) 和循环神经网络 (recurrent neural network, RNN), 在捕获短期时序依赖方面表现良好, 但在处理跨越长时序的结构性依赖时易出现信息遗忘等问题. 音乐中普遍存在跨越长时间段的复杂结构和依赖关系, 导致传统 RNN 类模型难以完整把握音乐全局特征, 限制了模型在长时序依赖要求较高的任务中的表现.

近年来, 基于 Transformer 架构的深度学习模型通过自注意力机制在捕获长序列依赖方面表现突出, 迅速成为 SMU 研究的主流. 该架构能够直接关注序列中任意位置的相关信息, 通过在海量未标注音乐上预训练, 再在下游任务中微调, 显著提升了模型精度[12,13]. MusicBERT[12]作为 SMU 方向的早期探索者之一, 成功将 BERT 架构[10]应用于大规模符号音乐数据集的预训练, 在旋律完成和伴奏建议等任务上展示了强大的潜力; MidiBERT[13]进一步针对符号音乐的特性进行了设计优化, 采用了掩码语言建模 (masked language modeling, MLM) 等预训练任务, 有力地证明了相较于传统 RNN 类模型, 大规模预训练模型在 SMU 任务上的显著优势. 这些工作充分验证了 Transformer 架构及其预训练方法在建模音乐序列方面的有效性.

为适应音乐的结构和层级特性, 研究者也对 Transformer 进行了针对性改进. 考虑到音乐中音符常以特定模式共存 (如和弦中的多个音符同时出现), 以及音乐结构具有层级性和相对时空关系[14], Li 等人[20]受旋转位置编码

(RoPE)[21]启发提出了旋转绝对-相对位置编码 (rotational absolute-relative, RoAR). RoAR 通过旋转嵌入机制灵活地表示音符间的相对空间关系, 以捕捉音乐序列中的层级结构和多维度特征, 尤其适用于理解结构复杂和动机不断演变的乐曲. 基于 RoAR 在层级依赖建模方面的卓越性能, CNN-Midiformer 将其集成为序列处理基线, 以进一步提升对音乐时序结构的建模精度.

然而, 基于 Transformer 架构的模型存在更深层的挑战: 模型仅通过序列数据学习进行建模, 难以有效学习音乐知识, 导致模型对音乐构成原理的理解相对有限, 限制了对音乐内在构造和深层语义的理解.

为了实现全面的理解能力, 研究者尝试引入非序列化的信息以弥补单一序列表征的不足. 部分研究尝试将乐谱图像作为新模态进行融合 (score images as a modality, SIM)[22], 旨在融合乐谱音符形状、谱表布局等视觉特征提供的非序列化信息和符号序列提供的精确时序信息, 但这种方法没有将结构化的音乐知识建模并融入模型表示中, 导致模型无法充分利用音乐知识来进一步提升性能.

当前 SMU 研究仍主要面临两项关键问题: 首先, 复杂音乐结构往往被简化为线性符号序列, 多维层级信息缺失; 其次, 乐理知识未有效融入模型表示, 导致对音乐内在构造原则的理解有限. 针对上述问题, 本文提出 CNN-Midiformer 模型, 通过将音乐知识转换为结构化表征, 利用卷积神经网络 (CNN) 提取关键局部特征, 与 Transformer 编码器捕获的全局语义特征通过高效交叉注意力融合, 显著提升了 SMU 基准下游任务的精度, 为 SMU 领域提供了一种新的研究思路.

2 融合音乐知识结构化表征的符号音乐理解模型

2.1 总框架

本文提出的 CNN-Midiformer 模型在符号音乐理解任务中引入音乐知识的结构化表征, 与符号序列信息协同建模, 在语义层面深化和扩展序列表示能力, 有效解决了传统 Transformer 模型在音乐表示单一化和乐理知识缺乏方面的局限性.

模型整体框架如图 2 所示, 由互补音乐特征提取模块、基于 RoAR[20]的序列信息捕获基线以及音乐知识自适应增强的特征融合模块这 3 大模块协同构成. 首先, 模型对 MIDI 文件进行结构化处理, 分别构建音乐知识 (音高转移矩阵和音长转移矩阵) 与音乐序列 (复合词) 两种结构化表征. 为有效捕获音高与音长之间的关联信息, 将两种矩阵堆叠形成堆叠特征矩阵. 随后, 互补音乐特征提取模块利用卷积神经网络 (CNN) 从堆叠特征矩阵提取局部结构特征, 获得深层音乐知识特征. 同时, 复合词序列通过 RoAR 基线处理, 捕获序列的深层语义信息.

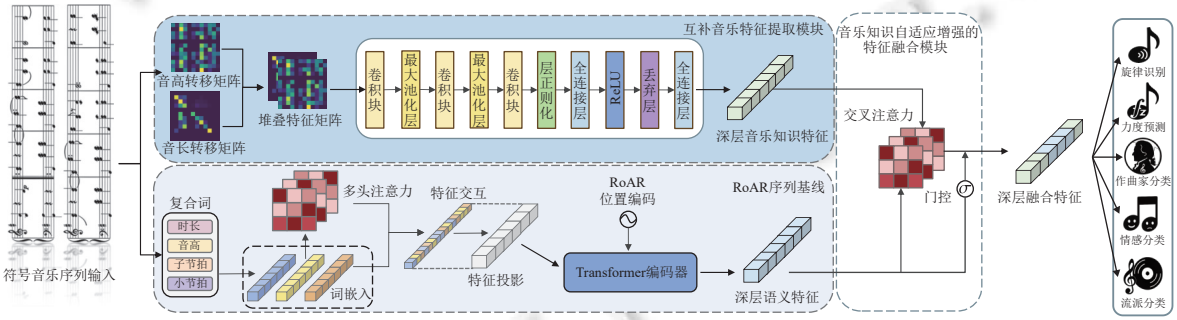


图 2 CNN-Midiformer: 融合音乐知识与序列信息的整体框架

最后, 音乐知识自适应增强的特征融合模块通过高效的交叉注意力机制, 根据 Transformer 编码器中每个符号的序列上下文信息, 自适应地将 CNN 提取的深层音乐知识特征融入序列表示之中, 确保音乐知识能在最相关的序列位置发挥作用.

2.2 互补音乐特征提取模块

音乐创作遵循旋律、和声与节奏等构造原则, 这些原则塑造了乐曲的独特层级结构与风格[14]. 音高转移矩阵 (PCTM) 与音长转移矩阵 (NLTM) 能够有效捕捉这些原则[16]. PCTM 通过统计各音高类别间的转移概率以刻画调

性特征与旋律逻辑, 例如在 C 大调中, C 到 G 的频繁转移反映了主音到属音的偏好; NLTM 则依据音符时值的转移规律建模节奏骨架与律动感, 例如 4 拍乐曲中 1/4 音符的接续频率揭示了稳定的节拍模式. 两者从统计角度构建了音乐在旋律与节奏维度上的内在结构与倾向, 为模型理解乐曲的理论基础与构造逻辑提供了有力支持.

基于该思考, 为解决传统 Transformer 模型在音乐序列建模中忽略音乐多维结构特征的问题, 本文提出了互补音乐特征提取模块. 该模块的设计基于以下两个观察: 一方面, 音乐知识特征 (音高转移矩阵与音长转移矩阵) 具备类似图像的二维空间结构, 包含丰富的局部空间关联模式, 蕴含乐曲中的旋律、节奏及风格等特征; 另一方面, CNN 在提取局部空间结构和层级特征方面表现优异, 与 Transformer 擅长捕获全局序列依赖关系形成有效互补. 利用 CNN 从音乐知识特征中提取局部结构特征, 并配合 Transformer 编码器捕获序列全局语义信息, 可以构成一种同时掌握音乐结构与音乐语义的完整理解框架.

构建音乐知识特征时, 首先将打击乐器 (如架子鼓) 轨道排除. 这是由于打击乐器的音高不具备明确的调性意义, 其节奏模式通常独立于旋律与和声轨道, 因此排除打击乐轨道可以使模型准确地捕获旋律、和声与节奏的核心结构特征. 音乐知识特征构建算法如算法 1 和算法 2 所示.

算法 1. PCTM 算法.

输入: MIDI 文件 I;

输出: 归一化音高类别转移矩阵 X, 尺寸 12×12.

1. $X \leftarrow zeros(12,12)$ // 初始化 12×12 的全 0 矩阵
2. **for** instrument ∈ I.instruments **do**
3. **if** instrument.is_drum then // 跳过打击乐器
4. continue
5. **end if**
6. $C \leftarrow \varnothing$ // 存放音高类别序列
7. **for** note ∈ instrument.notes **do**
8. $c \leftarrow$ note.pitch mod 12 // 映射到音高类别
9. append c to C
10. **end for**
11. **for** $t = 1$ to $|C|-1$ **do** // 统计相邻音符对
12. $i \leftarrow C[t-1]$
13. $j \leftarrow C[t]$
14. $X[i,j] \leftarrow X[i,j]+1$
15. **end for**
16. **end for**
17. **for** $i = 0$ to 11 **do** // 行归一化
18. $s \leftarrow \sum_{j=0}^{11} X[i,j]$
19. **if** $s \neq 0$ **then**
20. **for** $j = 0$ to 11 **do**
21. $X[i,j] \leftarrow X[i,j]/s$
22. **end for**
23. **end if**
24. **end for**
25. **return** X

算法 2. NLTM 算法.

输入: MIDI 文件 I; 时值类别集合 $D=\{d_0,\ldots,d_{11}\}$;
输出: 归一化时值转移矩阵 Y, 尺寸 12×12.

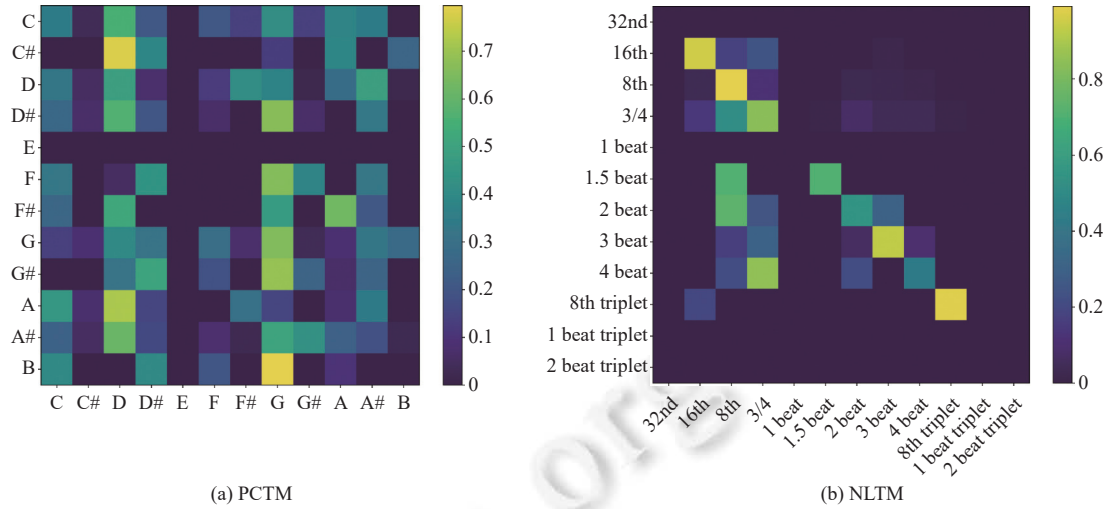
$\tau_{\text{beat}} \leftarrow I$.ticks_per_beat // 每拍的 ticks
$\tau \leftarrow [\tau_0,\ldots,\tau_{11}]$ // 由 τ_{beat} 推导的 12 类时值 ticks
$Y \leftarrow zeros(12,12)$ // 初始化 12×12 的全 0 矩阵
for instrument ∈ I.instruments **do**
  if instrument.is_drum **then** // 跳过打击乐器
    continue
  end if
  $\Lambda \leftarrow \varnothing$ // 存放时值类别序列
  for note ∈ instrument.notes **do**
    $\Delta \leftarrow$ note.end − note.start // 时值 (ticks)
    $k \leftarrow \text{argmin}_{p\in\{0,\ldots,11\}}|\Delta-\tau_p|$ // 匹配最近类别
    append k to Λ
  end for
  for $t=1$ to $|\Lambda|-1$ **do** // 统计相邻音符对
    $k \leftarrow \Lambda[t-1]$
    $l \leftarrow \Lambda[t]$
    $X[k,l] \leftarrow X[k,l]+1$
  end for
end for
for $k=0$ to 11 **do** // 行归一化
  $s \leftarrow \sum_{l=0}^{11} Y[k,l]$
  if $s \neq 0$ **then**
    for $l=0$ to 11 **do**
      $Y[k,l] \leftarrow Y[k,l]/s$
    end for
  end if
end for
return Y

通过算法 1 构建音高转移矩阵 (PCTM) 和算法 2 构建音长转移矩阵 (NLTM), 得到两个归一化的 12×12 矩阵, 它们共同组成了音乐知识特征. 这两种矩阵不依赖具体序列顺序, 从统计上有效地刻画了乐曲在音高和节奏维度上的内在结构与倾向性. 图 3 直观地展示了同一 MIDI 文件建模得到的 PCTM (图 3(a)) 和 NLTM (图 3(b)) 的可视化示例, 不同的颜色强度代表了不同的转移概率, 从中可以观察到该乐曲在音高和节奏上的特定规律.

随后, 将音乐知识特征堆叠形成二维特征图输入到互补音乐特征提取模块. 如图 4 所示, 模块通过多层的 3×3 卷积块操作, 捕捉输入特征中的局部空间特征, 卷积块操作可以表示为:

$$F_{i,j}^{l}=\sigma\left(BN\left(\sum_{m=0}^{K-1}\sum_{n=0}^{K-1}W_{m,n}^{l}\cdot X_{i+m,j+n}^{l-1}+b^{l}\right)\right) \tag{1}$$

其中, $F^l_{i,j}$ 表示第 l 层特征图在位置 (i,j) 的值, X^{l-1} 表示上一卷积块的输出, W^l 是卷积核权重, K 是卷积核大小, b^l 是偏置项, BN 是批正则化, σ 是 ReLU 激活函数.



(a) PCTM (b) NLTM

图 3 同一 MIDI 文件的 PCTM 和 NLTM

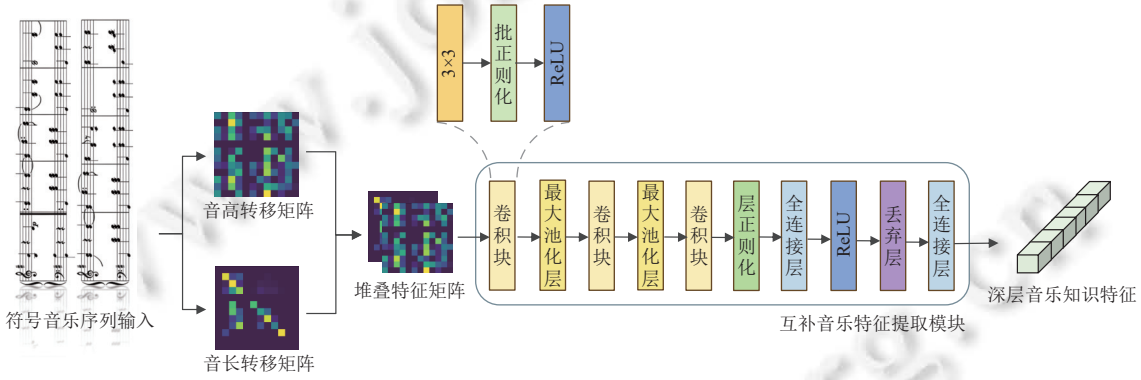


图 4 互补音乐特征提取模块

最大池化层级联卷积块, 用于降低特征图的空间维度, 提取具有代表性的抽象特征:

$$P^l_{i,j} = \max_{0\leqslant m,n<S} F^l_{S\cdot i+m,S\cdot j+n} \tag{2}$$

其中, $P^l_{i,j}$ 是池化后的特征, S 是池化窗口大小 (该模块的窗口大小 S=2).

为了稳定训练过程并加速模型收敛, 批正则化 (BN) 和层正则化 (LN) 被整合到网络中. 批正则化可表示为:

$$\hat{x}^l = \frac{x^l-\mu_B}{\sqrt{\sigma_B^2+\varepsilon}}\cdot\gamma+\beta \tag{3}$$

其中, μ_B 和 σ_B^2 是批次内的均值和方差, γ 和 β 是可学习的缩放和偏移参数, ε 是一个微小常数.

层正则化对卷积层的输出进行处理:

$$\hat{h} = \frac{h-\mu_h}{\sqrt{\sigma_h^2+\varepsilon}}\cdot\gamma+\beta \tag{4}$$

其中, μ_h 和 σ_h^2 是特征向量的均值和方差.

经过多层特征提取和正则化后, 网络通过全连接层将高维结构化特征输出为固定维度 (Transformer 编码器的

隐藏层维度, 本模型为 768) 的向量:

$$z = W_2 \cdot Dropout(\sigma(W_1 \cdot flatten(P^L) + b_1)) + b_2 \tag{5}$$

其中, P^L 是最后池化层输出的特征图, *flatten* 操作将其展平为一维向量, W_1、W_2、b_1、b_2 是全连接层的参数, σ 是 ReLU 激活函数, *Dropout* 操作用于随机丢弃训练参数, 避免模型过拟合.

最终, z 向量封装了音高和节奏的深层抽象表示, 用于与序列特征进一步融合.

2.3 音乐知识自适应增强的特征融合模块

音乐知识自适应增强的特征融合模块旨在将由 CNN 提取的结构化音乐知识与 Transformer 编码器所捕获的符号序列信息进行动态交互, 以实现语境相关的知识融合而非简单叠加. 该模块通过交叉注意力机制, 将音乐知识特征作为查询, Transformer 编码器隐藏层的序列表示作为键和值, 计算注意力权重并据此生成知识增强表示, 使模型能够根据每个符号在序列中的上下文动态提取最相关的音乐知识; 这一过程既避免了特征冗余, 又确保了知识对序列语义的精确指导.

根据 Transformer 编码器分层研究结果, 低层编码器关注句法信息, 高层侧重语义信息, 中间层则具有较强的任务灵活性[23]. 结合这一研究结论, 该模块选择 Transformer 编码器的第 2、5、8、11 层进行特征融合, 涵盖句法主导层 (低层)、语义过渡层 (中层)、任务专属层 (高层) 不同阶段, 在减少计算量的同时, 使音乐知识特征贯穿于模型的“句法-语义-决策”全流程.

图 5 展示了音乐知识自适应增强的特征融合模块架构. 设第 l 层的 Transformer 编码器隐藏状态为 $H^l = \{h_1^l, h_2^l, \ldots, h_{N+1}^l\}$, 其中 N 是序列长度, h_1^l 是特殊的 [CLS] 标记的表示, $h_2^l, \ldots, h_{N+1}^l$ 是音乐序列, z_{CNN} 为 CNN 提取的知识特征向量. 在选定的 Transformer 编码器层中, 首先对隐藏状态与知识向量分别进行线性变换:

$$\begin{cases} Q = W_Q \cdot z_{\text{CNN}} \\ K = W_K \cdot H^l \\ V = W_V \cdot H^l \end{cases} \tag{6}$$

其中, W_Q、W_K、W_V 是线性变换矩阵. 注意力权重和融合输出计算如下:

$$A = Softmax\left(\frac{Q \cdot K^{\mathrm{T}}}{\sqrt{d_K}}\right) \tag{7}$$

$$F = A \cdot V \tag{8}$$

其中, d_K 是注意力机制的缩放因子 (隐藏层大小为 768, d_K=768), F 是特征融合结果.

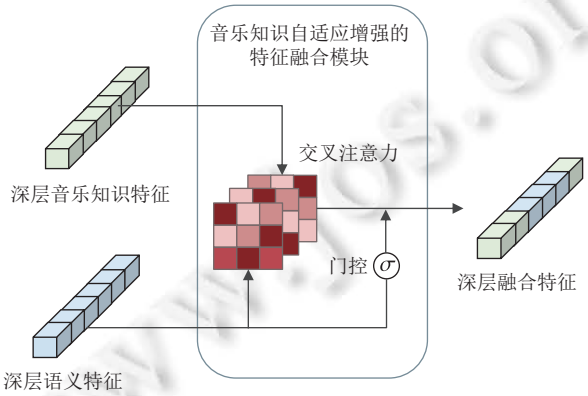


图 5 音乐知识自适应增强的特征融合模块

通过这种方式, 音乐知识特征能够关注序列表示中与其最相关的部分, 并根据这种相关性动态地从序列语境中提取信息, 生成融合了序列信息的知识增强表示.

为动态地调控原始序列表示与这个知识增强表示的结合比例, 模块引入了门控机制. 对序列中的每个位置 i, 门控值的计算方法如下:

$$G_i = \sigma(W_g \cdot [h_i^l : F_i] + b_g) \tag{9}$$

其中, $[h_i^l : F_i]$ 表示原始隐藏状态和融合特征的拼接, W_g 和 b_g 是可学习的参数, σ 是 Sigmoid 激活函数.

最终, 通过加权求和的方式将两者融合, 形成该 Transformer 编码器层的新隐藏状态:

$$\hat{h}_i^l = G_i \odot h_i^l + (1 - G_i) \odot F_i \tag{10}$$

其中, $\odot$ 表示元素级乘法.

这种逐层、动态、基于注意力和门控的融合策略, 使得音乐知识特征能够以序列上下文感知的方式自适应地在 Transformer 编码器处理序列信息的不同阶段施加影响, 从而引导模型在理解序列依赖的同时, 也能充分利用音乐的内在结构规律, 最终获得对符号音乐全面的理解.

2.4 训练对象

为确保本文提出的 CNN-Midiformer 模型有效结合音乐知识与符号序列信息, 并实现全面的音乐表示, 设计了一种适当的训练策略. 该训练策略能够驱动模型的所有组件协同优化, 尤其确保 CNN 提取的结构特征对符号音乐理解任务具有明确的贡献.

本文设计的训练策略结合了掩码语言建模 (MLM) 和因果语言建模 (causal language modeling, CLM) 两种目标, 通过端到端训练过程隐式优化卷积特征提取模块. 这种联合策略能够同时捕捉符号音乐序列中的双向上下文依赖和单向时序依赖, 对模型理解音乐进行具有关键作用. 同时, 通过整体训练过程, 确保 CNN 提取的音乐知识特征有效地服务于序列理解任务.

目标 1: 掩码语言建模 (MLM). 受到 BERT 训练方法的启发, MLM 目标通过让模型重建序列中被掩盖的符号来学习上下文相关的表示. 给定一个符号音乐序列 $S = \{s_1, s_2, \ldots, s_{N+1}\}$, 随机选择 15% 的符号进行掩码操作. 遵循标准的掩码协议, 其中 80% 被替换为特殊的 [MASK] 标记, 10% 被替换为词汇表中随机选择的其他符号, 剩余 10% 保持不变 (以减少预训练和微调阶段之间的差异). 模型需要利用被掩盖位置之前和之后的上下文信息来预测原始符号. MLM 的目标函数定义为:

$$L_{\text{MLM}}(\theta) = -\sum_{S_i \in M} \log P(S_i | S_{\setminus M}; \theta) \tag{11}$$

其中, M 代表被掩码符号索引的集合, $S_{\setminus M}$ 是带有掩码位置的输入序列, θ 代表模型的参数. 这种训练方式促使模型学习序列内丰富的语义表示, 捕捉音高、时长以及小节拍、子节拍等元素之间复杂的关系.

目标 2: 因果语言建模 (CLM). 为了补充 MLM 的双向理解能力, 本文同时采用了类似于 GPT 的单向生成目标. 在 CLM 任务中, 模型学习基于其之前的所有符号来预测当前位置的符号, 有效建模音乐从左到右的自然顺序. 对于序列 S 中的每个位置 i, 模型估计符号 S_i 基于前缀 $S_{<i} = \{s_1, s_2, \ldots, s_{i-1}\}$ 的条件概率. CLM 损失函数定义为序列中所有符号的负对数似然之和:

$$L_{\text{CLM}}(\theta) = -\sum_{i=1}^{N} \log P(S_i | S_{<i}; \theta) \tag{12}$$

该目标对于捕捉音乐的进行至关重要, 因为起奏时间、持续时间、力度变化等音符事件的时序关系是音乐连贯性和表现力细微差别的基础.

联合训练与知识特征优化: 模型并非孤立地训练这两个目标, 而是在训练过程中交替执行 MLM 和 CLM 任务. 在计算 MLM 和 CLM 损失并进行反向传播时, 梯度不仅会更新 Transformer 编码器的参数, 还会通过音乐知识自适应增强的特征融合模块回传到互补音乐特征提取模块. 虽然没有为互补音乐特征提取模块定义一个独立的损失函数, 但其参数会根据其提取的音乐知识特征对提升 MLM 和 CLM 任务性能的贡献程度而被优化. 这种端到端的训练机制确保了 CNN 学习到的结构化知识表示是与序列理解任务高度相关的, 并且能够有效增强 Transformer 编码器的表达能力. 总的训练损失定义为 MLM 和 CLM 损失之和:

$$L_{\text{total}} = L_{\text{CLM}} + L_{\text{MLM}} \tag{13}$$

通过整合这两个序列层面的目标任务, 并在整个模型中一致地进行优化, CNN-Midiformer 既能够从上下文中推断被掩盖的符号, 也能基于时序依赖预测未来符号, 使音乐知识特征中蕴含的结构化信息得到有效利用, 并进一步增强模型对音乐序列的综合理解能力.

3 实验结果与分析

3.1 数据集

本文在 6 个公开符号音乐理解数据集 Pop1K7、ASAP、POP909、Pianist8、EMOPIA 和 ADL 上进行了模型的预训练与下游任务微调, 各数据集的详细信息如表 1 所示.

表 1 实验所用数据集统计信息

数据集	任务	乐曲数 (pcs.)	总时长 (h)	平均音高 (MIDI No.)	平均音长 (beats)	小节音数 (notes/bar)	小节数量 (bars/pc.)
Pop1K7	预训练	1 748	108.8	64 (E4)	8.5	16.9	103.3
ASAP		65	3.5	63 (D4#)	2.9	23.0	95.9
POP909	旋律识别/力度预测	865	59.7	63 (D4#)	6.1	17.4	94.9
Pianist8	作曲家分类	411	31.9	63 (D4#)	9.6	17.0	108.9
EMOPIA	情感分类	1 078	12.0	61 (C4#)	10.0	17.9	14.8
ADL	流派分类	6 213	366.6	62 (D4)	2.7	16.1	104.9

- Pop1K7[24]: 包含 1 748 首流行钢琴曲的录音, 总时长约 108 h. 这些录音通过自动化流程被转录为 MIDI 文件.
- ASAP[25]: 包含 222 份数字乐谱和 1 068 个西方古典钢琴曲的 MIDI 演奏录音, 涵盖了 15 位不同作曲家的作品, 为模型提供了多样且高质量的古典音乐素材.
- POP909[26]: 包含 909 首由专业音乐家创作并演奏的流行歌曲钢琴演奏 MIDI. 该数据集完整保留旋律和力度信息, 特别适用于旋律识别和演奏速度预测等任务.
- Pianist8[13]: 包含由 8 位著名作曲家创作的 411 部钢琴作品, 其 MIDI 文件是使用 Kong 等人[27]提出的乐谱转录模型生成的, 该数据集主要用于作曲家风格分类等任务.
- EMOPIA[28]: 包含来自 387 首歌曲的 1 087 个钢琴片段, 每个片段均带有情感标签. 该数据集专为音乐情感分类任务设计, 为训练模型的情感识别能力提供丰富样本.
- ADL[29]: 包含 6 213 首 MIDI 文件, 分属 18 类不同的音乐风格, 覆盖广泛的音乐流派, 可用于音乐流派分类与风格分析任务.

3.2 训练设置

本文模型 CNN-Midiformer 基于 Hugging Face Transformers 库[30]实现, 并使用 PyTorch 框架进行训练. 在训练之前对每个 MIDI 文件均提取了音乐知识和符号序列的结构化表征. 具体的模型配置参数详见表 2.

预训练: 本文将每个数据集按照 85% 训练集和 15% 验证集的比例进行划分. 为了确保实验的可复现性, 下面将详细说明训练配置: 优化器选用 AdamW[31], 这是一种解耦了权重衰减与梯度计算的 Adam 变体, 已被证明能改善深度学习模型的泛化能力. 经过充分的超参数调优, 学习率设置为 2E–5, 该学习率在训练稳定性和收敛速度之间取得了较好的平衡. 权重衰减系数设置为 0.01, 以防止模型在音乐数据上发生过拟合. 最大序列长度设定为 512, 这个长度既考虑了数据集中乐曲片段的平均长度, 也平衡了计算资源消耗, 同时确保模型能够处理足够长的音乐上下文以捕捉音乐的层级结构. 预训练在 4 块 NVIDIA A800 GPU 上进行, 批次大小设为 12, 训练耗时约 23 h. 采用了早停策略: 若验证集准确率连续 30 个周期未提升, 则停止训练, 最大训练周期设为 500.

下游任务微调: 对于各项下游任务, 本文将数据按照 8:1:1 的比例划分为训练集、验证集和测试集. 微调同样在 4 块 NVIDIA A800 GPU 上进行, 学习率设置为 2E–5, 训练周期设为 20, 保存验证集准确率最高的模型.

表 2 模型超参数设置

超参数	值
最大序列长度	512
网络层数	12
隐藏层大小	768
注意力头数	12
权重衰减系数	0.01
优化器	AdamW
学习率	2E–5
批次大小	12
总参数量	113M

3.3 模型性能对比

本节在旋律识别、力度预测、作曲家分类、情感分类和流派分类这 5 项符号音乐理解基准任务上对比了 CNN-Midiformer 与当前多种先进模型的性能, 并分析了各模型在不同任务中表现提升的根本原因. 表 3 列出了包括非预训练的传统 RNN 模型、基于大规模预训练的 MidiBERT 与其改进版本以及引入改进位置编码的 RoAR 等模型在内的多种对照模型在上述任务上的准确率, 以 RoAR† (本文复现的基线模型) 为对比基线, 确保实验环境的一致性与结果的可比性. 表 3 中, 加粗表示该任务的最佳结果, 括号内的数值表示相对于基线 RoAR†的提升, ↑表示数值越高越好.

表 3 模型在下游任务上的性能对比 (%)

模型	词元分类		序列分类		
	旋律识别↑	力度预测↑	作曲家分类↑	情感分类↑	流派分类↑
MidiGPT[29]	—	—	—	61.88	—
RNN[13]	88.66	43.77	60.32	54.13	—
MidiBERT[13]	96.37	51.63	78.57	67.89	61.49
MT-MidiGPT[32]	—	—	—	66.95	—
MT-MidiBERT[32]	—	—	—	69.97	—
OM-MIDI-BERT[33]	97.87	53.42	75.40	68.81	—
RoAR[20]	97.59	53.73	80.95	76.15	—
RoAR†	97.05	53.21	79.37	75.23	64.49
CNN-Midiformer	**97.43 (0.38)**	**53.42 (0.21)**	**86.51 (7.14)**	**79.82 (4.59)**	**68.46 (3.97)**

在旋律识别与力度预测两项任务中, CNN-Midiformer 相较于 RoAR† 分别实现了 0.38 和 0.21 个百分点的提升. 尽管提升幅度有限, 但在高基线水平下依然具备显著意义. 旋律识别任务中, RoAR† 已达到 97.05% 的高准确率, 几乎贴近饱和状态; CNN-Midiformer 通过音乐知识特征与序列特征的融合, 虽未大幅突破该瓶颈, 却在细微旋律走向与跨拍依赖处理中展现了稳定的泛化能力. 在力度预测任务中, 模型对音符强弱动态的建模能力得到进一步提升. NLTM 对节奏律动的量化表示以及 CNN 对时值结构化特征的有效表征, 为模型提供了超出自注意力机制的额外力度感知, 从而更有效地捕捉力度随时间的细微变化.

在作曲家分类、情感分类和流派分类这 3 项任务中, 模型均取得了显著性能提升, 提升幅度分别达到了 7.14、4.59 和 3.97 个百分点. 这一现象表明, 结构化音乐知识在区分作曲家风格、情感表达和音乐流派方面具有重要作用. 在作曲家分类任务中, PCTM 有效刻画了不同创作者在旋律进行过程中的音高偏好, 这些偏好体现了作曲家的个性化写作习惯; 将 CNN 提取的 PCTM 特征与 Transformer 的全局语义信息融合后, 模型能够敏锐地捕获那些仅靠线性序列难以辨识的个性化风格标志[34]. 在情感分类任务中, NLTM 所体现的节奏密度与“激发度”等情感维度密切相关[35], 使模型能够获得更强的情感判别能力. 在流派分类任务中, CNN-Midiformer 在涵盖 18 个不同音乐流

派、类别分布更为复杂的 ADL 数据集上仍取得了 3.97 个百分点的性能提升, 展示出其在跨风格任务中的优异泛化能力. 总体而言, 通过互补音乐特征提取模块对节奏与律动的深层结构特征进行提炼, 与序列编码器融合后, 模型构建了对情感标签和流派标签的精准映射, 推动了分类精度的显著提升.

除上述任务外, CNN-Midiformer 在对比非预训练 RNN 模型时亦体现出压倒性优势. RNN 模型在旋律识别与力度预测上的准确率分别为 88.66% 和 43.77%, 远低于本文模型的 97.43% 和 53.42%, 凸显了大规模预训练与知识融合的重要性. MidiBERT 及其变体在多个基准任务中的表现也落后于 CNN-Midiformer, 说明在符号音乐理解领域, 仅依赖序列预训练而缺乏结构化乐理知识融合的模型在高难度任务中仍存在性能瓶颈.

综上所述, CNN-Midiformer 将通过数学建模的音乐知识特征与强大的自注意力序列编码器有机融合, 在 5 项任务中均取得了对比模型无法比拟的性能优势, 使模型能够同时兼顾局部结构与全局依赖. 在作曲家分类、音乐情感分类和流派分类这 3 项高度依赖结构信息的任务上, 知识特征的有效应用显著扩展了模型的判别能力, 验证了模型在不同音乐风格和任务场景中的通用性.

3.4 消融实验

为量化各模块与训练策略对 CNN-Midiformer 性能贡献, 本文进行系列消融实验. 结果如表 4 所示, 互补音乐特征提取模块、RoAR 位置编码、预训练与端到端微调策略在不同层面共同决定了模型的卓越性能, 并结合 CNN 核尺寸对性能的影响分析 (表 5) 进行了系统验证. 表 4 和表 5 中, 加粗表示该任务的最佳结果, ↑表示数值越高越好.

表 4 模型消融实验结果汇总 (%)

模型	词元分类		序列分类		
	旋律识别↑	力度预测↑	作曲家分类↑	情感分类↑	流派分类↑
RoAR†	97.05	53.21	79.37	75.23	64.49
Full model	**97.43**	53.42	**86.51**	**79.82**	**68.46**
fine-tune w/o CNN	97.17	53.10	84.13	77.06	66.28
w/o CNN	97.05	53.21	79.37	75.23	64.49
w/o RoAR	97.21	**53.73**	80.95	76.15	66.82
w/o fine-tune	91.30	44.77	69.05	65.14	—
w/o pre-train	94.32	45.79	58.73	56.67	—

表 5 互补特征提取模块核尺寸对性能的影响 (%)

模型	词元分类		序列分类		
	旋律识别↑	力度预测↑	作曲家分类↑	情感分类↑	流派分类↑
RoAR†	97.05	53.21	79.37	75.23	64.49
Kernel 3×3 (Full model)	97.43	**53.42**	**86.51**	**79.82**	**68.46**
fine-tune w/o CNN	97.17	53.10	84.13	77.06	66.28
Kernel 5×5	**97.46**	53.21	81.75	71.56	68.15
Kernel 7×7	97.09	53.12	83.33	66.97	64.41

当完全移除互补音乐特征提取模块 (w/o CNN) 时, 模型在所有 5 项任务上的表现均回落至 RoAR 基线水平, 旋律识别与力度预测准确率分别降至 97.05% 与 53.21%, 作曲家分类、情感分类和流派分类降至 79.37%、75.23% 和 64.49%. 这一现象表明, CNN 对 PCTM 与 NLTM 所体现的局部结构信息的提炼在提升整体性能中发挥了关键作用; 缺失音乐知识的融合后, 模型退化为仅依赖 Transformer 序列表示, 导致性能全面下降, 进一步验证了互补音乐特征提取模块对整体性能提升的必要性.

仅在预训练阶段保留互补特征提取模块而在下游任务微调时移除 (fine-tune w/o CNN), 模型性能虽较完全移除时有所提升, 但仍低于完整模型. 在旋律识别与力度预测任务中, 准确率分别降至 97.17% 与 53.10%, 而作曲家分类、情感分类和流派分类则降至 84.13%、77.06% 和 66.28%. 这一结果表明, 预训练阶段学习的音乐知识特征对后续任务具有持续影响, 但如果不在微调阶段继续利用这些结构化特征, 其效用将受到削弱, 体现了端到端联合

优化的重要性.

移除 RoAR 位置编码 (w/o RoAR) 后, 模型在旋律识别与力度预测任务中准确率分别为 97.21% 与 53.73%, 作曲家分类、情感分类和流派分类则为 80.95%、76.15% 和 66.82%, 整体性能较完整模型有所下降, 却依然超越 RoAR† 基线. 这说明 RoAR 在捕捉长距离时序依赖方面确有助益, 但结构化音乐知识特征的有效融合同样能够在没有该位置编码的情况下取得显著增益, 展现了二者在补充和强化序列信息方面的互补特性.

在完全不进行预训练 (w/o pre-train) 的设置下, 各项任务的准确率出现显著滑坡: 旋律识别准确率下降至 91.30%, 力度预测性能下滑至 44.77%, 作曲家分类与情感分类任务的准确率也分别大幅跌至 69.05% 和 65.14%. 这一结果凸显了预训练环节在学习通用音乐表示中的基础作用, 缺乏大规模数据的预训练, 模型难以形成对音乐序列及其结构化特征的充分表征能力, 因而在多项下游任务中表现明显受限.

若仅微调分类头而冻结其余网络参数 (w/o fine-tune), 则性能也出现明显下滑, 旋律识别与力度预测分别降至 94.32% 与 45.79%, 作曲家分类与情感分类降至 58.73% 与 56.67%. 这一现象表明, 全模型微调对于适应具体下游任务至关重要, 仅调整头部分类层无法充分利用预训练得到的序列与知识特征.

通过对上述消融实验的综合分析可以得到, 互补音乐特征提取模块、RoAR 位置编码、预训练以及端到端微调策略在不同层面共同支撑了 CNN-Midiformer 的卓越性能: 互补音乐特征提取模块与 RoAR 分别负责建模局部结构和捕获全局依赖, 是模型性能提升的重要组成部分; 预训练阶段所学得的通用表示为下游任务奠定了稳健基础; 而端到端的联合微调确保了结构化知识与序列表征在具体任务中的有效融合, 最终实现了模型性能的全面提升.

为进一步验证 CNN 参数配置对模型性能的影响, 本文对不同卷积核尺寸进行了对比实验. 结果显示, 3×3 卷积核在整体表现上更为稳定: 在力度预测任务中达到 53.42% 的准确率, 相较于 5×5 核的 53.21% 和 7×7 核的 53.12% 均有提升. 值得注意的是, 尽管 5×5 卷积核在旋律识别任务上略占优势, 但在作曲家分类和情感分类等更具挑战性的任务中, 3×3 卷积核表现更佳, 准确率分别达到 86.51% 和 79.82%. 这一结果表明, 合适的卷积核尺寸对平衡多类型特征的表征具有重要意义: 小卷积核能够更有效地捕获局部音乐结构, 缓解大卷积核可能引入的过度平滑问题, 从而保留更丰富的细粒度信息. 在流派分类任务中, 3×3 卷积核同样取得了最高准确率 (68.46%), 进一步验证了该参数配置在不同任务上的有效性与普适性.

表 6 展示的门控值统计进一步阐释了特征融合过程中不同层级对音乐知识融合的敏感度. 第 2 层与第 11 层的平均门控值分别达到 0.8093 与 0.9422, 显著高于中间层 (第 5 层 0.6940, 第 8 层 0.6996). 这一分层表现与 Transformer 编码器分层研究[23]的结论高度一致: 底层更侧重句法结构, 高层则贴近具体任务需求, 这表明, 在句法主导阶段与决策输出阶段, 编码器分别处于句法结构构建与全局任务信息充分聚合的状态, 仅需适量引入音乐知识即可与原有表示形成有效协同; 中间层则处于由句法向语义过渡、持续抽象任务要素的过程, 更依赖音乐知识的深度融合, 以补充结构化的乐理知识, 增强中间表示的完备性.

表 6 各层门控值统计

隐藏层编号	门控均值
2	0.8093
5	0.6940
8	0.6996
11	0.9422

综上所述, 消融实验定量地分析验证了 CNN-Midiformer 中各组件与训练策略的设计必要性, 为进一步优化符号音乐理解模型提供了清晰的理论支撑. 尽管模型在多项任务中取得了显著进展, 但仍存在音符密度偏差与标签噪声场景下的性能瓶颈, 详细的失败案例分析与可视化结果参见附录 A.

4 总 结

本文提出的 CNN-Midiformer 模型通过融合结构化音乐知识与符号序列信息, 实现了对结构化乐理知识与符号序列表示的深度协同建模: 前者利用卷积神经网络提取旋律与节奏的局部构造规律, 后者基于交叉注意力机制

在序列语境中动态融合知识特征, 从而同时兼顾多层次结构与长距离依赖, 有效缓解了纯序列模型忽略音乐层级规律且缺乏乐理知识的问题. 在 6 个公开符号音乐理解数据集 Pop1K7、ASAP、POP909、Pianist8、EMOPIA 和 ADL 上的实验结果表明, 该模型在旋律识别、力度预测、作曲家分类、情感分类和流派分类这 5 项符号音乐理解基准任务上均优于基线模型, 尤其在作曲家分类、情感分类和流派分类任务上分别取得 7.14、4.59 和 3.97 个百分点的显著性能提升. 消融实验进一步验证了音乐知识特征对深层音乐结构与语义理解的关键作用. 本文提出的方法为符号音乐理解领域提供了新的研究视角, 也为音乐生成、实时情感识别和多模态音乐分析提供了技术支撑.

References

[1] Salamon J, Gómez E, Ellis DPW, Richard G. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. IEEE Signal Processing Magazine, 2014, 31(2): 118–134. [doi: 10.1109/MSP.2013.2271648]

[2] Simonetta F, Cancino-Chacón CE, Ntalampiras S, Widmer G. A convolutional approach to melody line identification in symbolic scores. In: Proc. of the 20th Int'l Society for Music Information Retrieval Conf. Delft: ISMIR, 2019. 924–931. [doi: 10.5281/zenodo.3527965]

[3] Jeong D, Kwon T, Kim Y, Lee K, Nam J. VirtuosoNet: A hierarchical RNN-based system for modeling expressive piano performance. In: Proc. of the 20th Int'l Society for Music Information Retrieval Conf. Delft: ISMIR, 2019. 908–915. [doi: 10.5281/zenodo.3527962]

[4] Jeong D, Kwon T, Kim Y, Nam J. Graph neural network for music score data and modeling expressive piano performance. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 3060–3070.

[5] Tsai T, Ji K. Composer style classification of piano sheet music images using language model pretraining. In: Proc. of the 21st Int'l Society for Music Information Retrieval Conf. Montreal: ISMIR, 2020. 176–183. [doi: 10.5281/zenodo.4245398]

[6] Kim S, Lee H, Park S, Lee J, Choi K. Deep composer classification using symbolic representation. arXiv:2010.00823, 2020.

[7] Grekow J, Raś ZW. Detecting emotions in classical music from MIDI files. In: Proc. of the 18th Int'l Symp. on Foundations of Intelligent Systems. Prague: Springer, 2009. 261–270. [doi: 10.1007/978-3-642-04125-9_29]

[8] Panda R, Malheiro R, Paiva RP. Musical texture and expressivity features for music emotion recognition. In: Proc. of the 19th Int'l Society for Music Information Retrieval Conf. Paris: ISMIR, 2018. 383–391. [doi: 10.5281/zenodo.1492431]

[9] Good M. MusicXML: An Internet-friendly format for sheet music. 2001. http://michaelgood.info/publications/music/musicxml-an-internet-friendly-format-for-sheet-music/

[10] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[11] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 2020, 30(4): 681–694. [doi: 10.1007/s11023-020-09548-1]

[12] Zeng ML, Tan X, Wang R, Ju ZQ, Qin T, Liu TY. MusicBERT: Symbolic music understanding with large-scale pre-training. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. ACL, 2021. 791–800. [doi: 10.18653/v1/2021.findings-acl.70]

[13] Chou YH, Chen IC, Ching J, Chang CJ, Yang YH. MidiBERT-Piano: Large-scale pre-training for symbolic music classification tasks. Journal of Creative Music Systems, 2024, 8(1): 1–19. [doi: 10.5920/jcms.1064]

[14] de Berardinis J, Vamvakaris M, Cangelosi A, Coutinho E. Unveiling the hierarchical structure of music by multi-resolution community detection. Trans. of the Int'l Society for Music Information Retrieval, 2020, 3(1): 82–97. [doi: 10.5334/tismir.41]

[15] Conklin D, Witten IH. Multiple viewpoint systems for music prediction. Journal of New Music Research, 1995, 24(1): 51–73. [doi: 10.1080/09298219508570672]

[16] Yang LC, Lerch A. On the evaluation of generative models in music. Neural Computing and Applications, 2020, 32(9): 4773–4784. [doi: 10.1007/s00521-018-3849-7]

[17] Le DVT, Bigo L, Herremans D, Keller M. Natural language processing methods for symbolic music generation and information retrieval: A survey. ACM Computing Surveys, 2025, 57(7): 175. [doi: 10.1145/3714457]

[18] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[19] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.

[20] Li ZC, Gong RH, Chen YN, Su KH. Fine-grained position helps memorizing more, a novel music compound Transformer model with feature interaction fusion. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 5203–5212. [doi: 10.1609/aaai.v37i4.25650]

[21] Su JL, Ahmed M, Lu Y, Pan SF, Bo W, Liu YF. RoFormer: Enhanced Transformer with rotary position embedding. Neurocomputing,

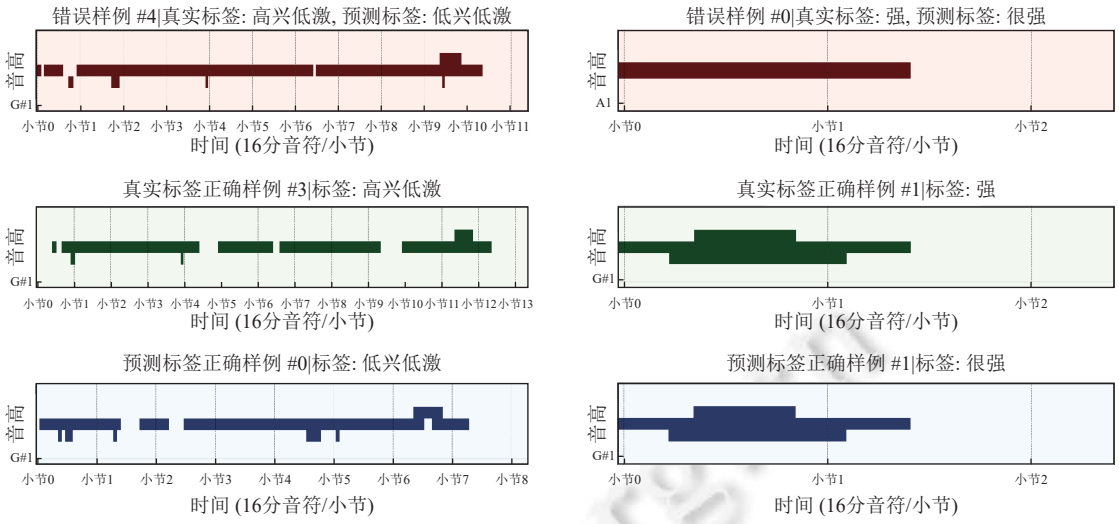
2024, 568: 127063. [doi: 10.1016/j.neucom.2023.127063]
[22] Qin Y, Xie HM, Ding SX, Li YJ, Tan BY, Ye MC. Score images as a modality: Enhancing symbolic music understanding through large-scale multimodal pre-training. Sensors, 2024, 24(15): 5017. [doi: 10.3390/s24155017]
[23] Clark K, Khandelwal U, Levy O, Manning CD. What does BERT look at? An analysis of BERT's attention. In: Proc. of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence: ACL, 2019. 276–286. [doi: 10.18653/v1/W19-4828]
[24] Hsiao WY, Liu JY, Yeh YC, Yang YH. Compound word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021, 35(1): 178–186. [doi: 10.1609/aaai.v35i1.16091]
[25] Foscarin F, McLeod A, Rigaux P, Jacquemard F, Sakai M. ASAP: A dataset of aligned scores and performances for piano transcription. In: Proc. of the 21st Int'l Society for Music Information Retrieval Conf. Montreal: ISMIR, 2020. 534–541. [doi: 10.5281/zenodo.4245490]
[26] Wang ZY, Chen K, Jiang JY, Zhang YY, Xu MR, Dai SQ, Xia G. POP909: A pop-song dataset for music arrangement generation. In: Proc. of the 21st Int'l Society for Music Information Retrieval Conf. Montreal: ISMIR, 2020. 38–45. [doi: 10.5281/zenodo.4245366]
[27] Kong QQ, Li BC, Song XC, Wan Y, Wang YX. High-resolution piano transcription with pedals by regressing onset and offset times. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 3707–3717. [doi: 10.1109/TASLP.2021.3121991]
[28] Hung HT, Ching J, Doh S, Kim N, Nam J, Yang YH. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In: Proc. of the 22nd Int'l Society for Music Information Retrieval Conf. ISMIR, 2021. 318–325. [doi: 10.5281/zenodo.5624518]
[29] Ferreira L, Lelis LHS, Whitehead J. Computer-generated music for tabletop role-playing games. In: Proc. of the 16th AAAI Conf. on Artificial Intelligence and Interactive Digital Entertainment. AAAI, 2020. 59–65. [doi: 10.1609/aiide.v16i1.7408]
[30] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu CW, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A. Transformers: State-of-the-art natural language processing. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. ACL, 2020. 38–45. [doi: 10.18653/v1/2020.emnlp-demos.6]
[31] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101, 2019.
[32] Qiu JB, Chen CLP, Zhang T. A novel multi-task learning method for symbolic music emotion recognition. arXiv:2201.05782, 2022.
[33] Liu SK, Xu HG, Xu K. An optimized method for large-scale pre-training in symbolic music. In: Proc. of the 16th IEEE Int'l Conf. on Anti-counterfeiting, Security, and Identification. Xiamen: IEEE, 2022. 105–109. [doi: 10.1109/ASID56930.2022.9995766]
[34] Yang D, Tsai TJ. Composer classification with cross-modal transfer learning and musically-informed augmentation. In: Proc. of the 22nd Int'l Society for Music Information Retrieval Conf. ISMIR, 2021. 802–809.
[35] Fernández-Sotos A, Fernández-Caballero A, Latorre JM. Influence of tempo and rhythmic unit in musical emotion regulation. Frontiers in Computational Neuroscience, 2016, 10: 80. [doi: 10.3389/fncom.2016.00080]

附录 A

A.1 音符密度偏差

在情感分类任务的典型失败案例中 (见图 A1(a)), CNN-Midiformer 出现了系统性的误判现象. 该样本在音高分布、时值均值及和弦复杂度等核心特征上均符合"高兴低激 (HALV)"类别的统计特征, 但模型却将其错误归类为"低兴低激 (LALV)". 深入分析发现, 关键问题在于音符密度的异常偏离: 该片段的音符密度达到 3.2 notes/beat, 相比同类 HALV 样本的均值高出约 15%, 并与 LALV 样本的密度分布区间产生显著重叠.

这一现象的根本原因在于互补音乐特征提取模块中音长转移矩阵 (NLTM) 对节奏律动变化的高度敏感性. 当 NLTM 检测到异常的高密度节奏结构时, CNN 层能够提炼出与 LALV 训练样本相似的密集节奏特征模式. 在特征融合阶段, 门控机制会在中间层自动调整各通道的权重分配, 显著提升节奏特征通道的影响力, 导致模型对情绪"激发度"维度的判断向 LALV 类别偏移. 消融实验结果进一步证实了这一机制: 在移除互补特征提取模块 (w/o CNN) 的配置下, 模型性能虽然回落至 RoAR 基线水平, 但不再对密度异常样本表现出过度敏感, 这表明该模块在提升整体性能的同时, 也成为此类误判的直接触发因素.



(a) 任务: 情感分类 (b) 任务: 力度预测

图 A1 情感分类和力度预测失败案例

A.2 标签噪声

在力度预测任务的失败案例中 (图 A1(b)), 模型的错误预测并非源于特征抽取能力的不足, 而是训练数据中存在的标签不一致问题. 具体表现为两条完全相同的音符序列在数据集中先后被人工标注为 f (强) 和 ff (很强), 形成了自相矛盾的监督信号. 这种标签冲突在模型训练过程中产生了复杂的负面影响: 模型在预训练阶段通过大规模数据建立了相对稳定的音高-力度映射关系, 但在下游任务微调时同时接收到相互矛盾的标签信息, 导致端到端优化过程无法收敛到唯一的决策边界.

在推理阶段, 模型面对这类边界模糊的样本时便倾向于在相邻类别间产生不确定性摇摆. 尽管互补特征提取模块的 PCTM 经 CNN 提炼后能够正确捕获音高子结构信息, RoAR 位置编码也有效传递了全局时序依赖关系, 但当监督信号本身存在不一致时, 即使互补音乐特征提取模块与序列编码器协同工作, 模型仍难以给出稳定一致的预测结果. 值得注意的是, 类似的标签不一致现象不仅出现在力度预测任务中, 在旋律识别任务中也同样存在: 当训练数据中的标注信号本身含有噪声时, 即便完整模型具备充分的学习能力, 也无法在相互矛盾的标签约束之间找到唯一最优解.

综上所述, 音符密度偏差和标签噪声揭示了模型在处理异常数据分布和不一致标注时的明确局限性, 凸显了高精度数据集对于提升模型性能的必要性. 未来的研究应侧重于开发能够降低这些局限性的方法, 例如改进特征提取模块以减少对异常数据的敏感性, 以及采用数据清洗和标签校正技术以提高数据集的标注质量.

作者简介

黄恒焱, 本科生, CCF 学生会员, 主要研究领域为音频信号处理, 音乐人工智能, 自然语言处理.

邹逸, 博士生, 主要研究领域为音乐人工智能, 自然语言处理.

时乐轩, 本科生, 主要研究领域为数字音乐生态研究, 新技术与音乐传播的融合, 中华民族音乐的当代传播与国际化路径.

程皓楠, 博士, 研究员, 主要研究领域为音频信号处理, 有声文化计算.

叶龙, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为多模态学习, 音视频生成与鉴伪.