郭圣濠, 蔡瑾, 晏峻峰. 基于异构图注意力网络的多模态音乐情感分类[J]. 智能计算机与应用,2026,16(3):58-66. DOI:10.20169/j.issn.2095-2163.26030101

基于异构图注意力网络的多模态音乐情感分类

郭圣濠[1,2], 蔡 瑾[1,2], 晏峻峰[1,2]

(1 湖南中医药大学 信息科学与工程学院, 长沙 410208; 2 湖南省智慧中医工程技术研究中心, 长沙 410208)

摘 要: 随着数字音乐规模的爆炸式增长,音乐情感识别(Music Emotion Recognition,MER)已经成为音乐信息检索、个性化推荐和情感计算领域的核心任务。传统单模态方法难以全面捕捉音乐的多维情感表达,并且现有多模态融合策略多采用简单拼接或 CNN-LSTM 结构,无法有效建模音频与歌词之间异步、异构的长程互补关系,导致分类准确率存在瓶颈。本文提出的 SongGATNet,是一种基于图注意力网络(Graph Attention Network,GAT)的音频歌词多模态融合模型。该模型将每首歌曲的多个 10 s 音频片段与歌词片段显式建模为异构图,通过跨模态全连接与同模态全连接实现细粒度特征交互,并且采用双层 GAT 编码器进行注意力机制下的跨模态信息融合,并结合混合全局池化(均值+最大池化)生成歌曲级表征,最后接入强化分类头(MLP+BatchNorm+ELU+Dropout)。实验在 NJU-MusicMood 数据集(共 765 首歌曲,共分成 4 类:Sad、Angry、Happy、Relaxed,500 首用于训练,265 首用于测试)上进行。SongGATNet 在测试集上取得准确率 0.94,宏平均 $F1$ 分数 0.94,显著优于基线模型。较 CNN-LSTM 提升 6 个百分点,较 Concat+MLP 提升 6 个百分点。消融实验进一步验证了异构图结构、跨模态边以及混合池化的关键贡献。

关键词: 音乐情感分类; 多模态学习; GAT

中图分类号: TP183 **文献标志码**: A **文章编号**: 2095-2163(2026)03-0058-09

Multimodal music emotion classification based on heterogeneous graph attention network

GUO Shenghao[1,2], CAI Jin[1,2], YAN Junfeng[1,2]

(1 School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China;
2 Hunan AI TCM Lab, Changsha 410208, China)

Abstract: With the explosive growth of digital music scale, Music Emotion Recognition (MER) has become a core task in the fields of music information retrieval, personalized recommendation, and affective computing. Traditional unimodal methods struggle to comprehensively capture the multi-dimensional emotional expression of music, while existing multimodal fusion strategies mostly rely on simple concatenation or CNN-LSTM architectures, which fail to effectively model the asynchronous and heterogeneous long-range complementary relationships between audio and lyrics, resulting in bottlenecks in classification accuracy. This paper proposes SongGATNet, an audio-lyrics multimodal fusion model based on Graph Attention Network (GAT). The model explicitly constructs a heterogeneous graph for multiple 10-second audio segments and lyric segments of each song, achieves fine-grained feature interaction through cross-modal and intra-modal fully-connected edges, and employs a two-layer GAT encoder for attention-based cross-modal information fusion. A hybrid global pooling strategy (mean + max pooling) is then applied to generate song-level representations, followed by a reinforced classification head (MLP + BatchNorm + ELU + Dropout). Experiments are conducted on the NJU-MusicMood dataset, comprising 765 songs with four emotion categories: Sad, Angry, Happy, and Relaxed. The dataset is split at the song level, with 500 songs for training and 265 for testing. SongGATNet achieves an accuracy of 0.94 and a macro-averaged $F1$ - score of 0.94 on the test set, significantly outperforming baseline models. This represents a 6-percentage-point improvement over CNN-LSTM and a 6-percentage-point improvement over Concat+MLP. Ablation studies further validate the critical contributions of the heterogeneous graph structure, cross-modal edges, and hybrid pooling mechanism.

Key words: Music Emotion Classification; multimodal learning; Graph Attention Network(GAT)

基金项目: 湖南省教育厅科学研究重点项目(23A312)。

作者简介: 郭圣濠(2000—),男,硕士研究生,主要研究方向:多模态融合,音乐情感分类; 蔡 瑾(1984—),女,博士研究生,主要研究方向:多模态融合,音乐情感分类。

通信作者: 晏峻峰(1965—),女,教授,博士生导师,主要研究方向:大数据与计算中医学。Email:junfengyan@hnucm.edu.cn。

收稿日期: 2026-03-01

0 引　言

音乐情感识别（Music Emotion Recognition，MER）是指利用计算机技术提取和分析音乐特征，在音乐特征与情感空间之间建立映射关系，从而识别音乐所表达情感的过程[1]。这一领域是音乐心理学、音频信号处理和自然语言处理等多学科的交叉研究[2-3]，目前 MER 在音乐信息检索、个性化推荐及情感计算等领域具有广泛应用前景。然而，传统依赖人工标注的方式不仅效率低下，而且不同人对同一首歌曲的情感感知往往存在差异，导致标注质量难以保证[4]，难以满足海量音乐的情感标注需求。在此背景下，音乐情感自动识别逐渐成为研究热点，现有研究主要集中于单模态建模，包括基于歌词文本的语义情感分类方法和基于音频信号的时频特征分析方法。然而，单一模态难以完整刻画音乐情感的多维表达：音频更侧重感知层面的情绪唤起，而歌词更偏向语义层面的情感表达。两者都存在明显局限，都存在着严重的分类准确率的瓶颈问题[5]，因此，近年来多模态融合方法逐渐成为提升 MER 性能的重要方向。基于此，本文聚焦音频与歌词的多模态融合，提出一种基于图注意力网络的多模态音乐情感分类方法 SongGATNet，旨在进一步提升音乐情感分类的准确率与效率。

本研究首先采用 Audio Spectrogram Transformer（AST）模型对音频模态进行特征提取：将每首歌曲切分为 10 s 音频片段，转换为 log-Mel 频谱图后输入预训练 AST 模型，附加 MLP 分类头实现四分类任务。同时，对于歌词模态，采用 BERT+CNN 模型：BERT 提取语义表征，叠加多核 CNN（核大小 3、4、5）捕获局部情感特征，测试阶段通过软投票聚合得到歌曲级预测。这些单模态方法为后续多模态融合提供了高质量的特征基础。

为了更有效地捕捉音频和歌词之间的深层关系，本文引入图神经网络对音频和歌词特征进行融合，采用图注意力网络（Graph Attention Network，GAT）实现特征间的选择性信息传递与融合，将音频片段与歌词片段建模为异构图，利用双层 GAT 实现注意力驱动的细粒度跨模态交互。该方法不仅保留了各模态的原始丰富特征，还通过图结构动态适应变长片段，无需固定长度填充，从而能够提升跨模态任务中的音乐情感识别能力。

本模型的创新点主要包括：

（1）提出一种面向异步多模态序列的统一图建模方法，实现跨模态信息的结构化表达。将音频歌词片段显式建模为异构图结构，将每首歌曲的多个 10 s 音频片段与歌词片段构建为一个无向异构图 $G=(V,E)$，其中节点 V 包含所有音频节点和歌词节点，边 E 同时包含跨模态全连接和同模态全连接。这种图建模方式实现了音频与歌词在片段级别的细粒度显式交互，既保留了各模态的原始丰富特征，又通过图结构动态适应不同歌曲的变长片段。

（2）提出双层 GAT 编码器实现注意力驱动的跨模态深度融合，在异构图基础上，采用双层 Graph Attention Network（GAT）编码器，通过多头自注意力机制动态计算邻居节点的加权聚合，实现音频与歌词之间长程依赖关系的精确建模。双层设计既能捕捉局部片段交互，又能建模全局情感关联，具有更强的可解释性和鲁棒性。

（3）基于预训练 AST 模型和 BERT+CNN 模型进行音频和歌词的特征提取，分别利用其强大的特征提取能力，提高多模态融合的精度。

综上所述，本文构建的 GAT 多模态音乐情感分类模型在保持结构简洁的同时，充分融合了音频和歌词的关键信息，为多模态音乐情感分类提供了更具实效的解决方案。

1 相关工作

1.1 音频分类

音频分类是音乐情感识别的重要基础。早期研究以机器学习方法为主，K 近邻（KNN）与支持向量机（SVM）最为常见。KNN 通过计算样本间距离进行决策，直观但易受 K 值与训练集分布影响。纪正飚等学者[6]引入模糊隶属度改进 KNN，提升了语音与视频情感识别效果。曹智贤[7]结合 K 近邻回归与 SVM，但在高可变性维度上取得效果有限。SVM 在高维特征处理中优势明显，泛化能力强[8]，能有效抑制过拟合。例如，魏华珍等学者[9]发现频谱、幅度谱、相位谱与时域特征融合效果最佳。王秀等学者[10]融合特征差异与 SVM 投票机制，提升了分类性能。此外，朴素贝叶斯方法在高维音频数据中分类效率高，在多类音乐情感识别中表现良好[11]。随着数据规模扩大，深度学习逐渐成为主流。CNN 在空间特征提取方面表现突出，结合迁移学习[12]与 Mel 谱图[13-14]，广泛应用于音乐情感分类。

近年来，复合模型成为研究热点。陈长风[15]构建了基于卷积长短期记忆网络（CNN-LSTM）的音乐情感模型，Tang 等学者[16]提出深度学习与广泛学

习混合框架,提升适应性与效率[17]。

总体而言,深度学习凭借强大特征提取能力成为音频情感分类的主导范式,但其在捕捉长程全局依赖方面仍存在局限。

1.2 歌词分类

歌词作为音乐作品的语义载体,蕴含丰富情感信息,在情感分析领域具有显著意义与应用价值。相较于音频信号,仅依靠歌词的情感分类研究起步较晚。

早期研究多基于传统文本表示方法,Chen 等学者[18]率先提出基于歌词的单模态情感识别任务,利用向量空间模型评估压力水平,但文本表征存在局限。Xia 等学者[19]构建情感向量空间模型,仅凭歌词即获得较优表现,验证了歌词在情感识别中的潜力。目前,面向歌词的音乐情感识别已成为情感计算的重要分支。陈若涵等学者与 Yang 等学者[20]从不同视角开展研究:前者比较 GMM、SVM 及 KNNR 等分类器的适用性差异;后者生成 182 维心理特征向量,借助决策树实现情感判别。两者共同推动了歌词情感分析方法的多样化。

伴随深度学习发展,神经网络在歌词处理中优势明显。杜常辉[21]在歌词配图任务中采用 LSTM 建模,通过门控机制有效捕捉长期依赖,性能显著优于传统算法,为歌词情感分析提供了更强大的建模工具。

由此可见,歌词情感分析正逐步向更高效、更精细的方向演进。然而,由于歌词情感表达的稀疏性,部分类别识别效果仍不理想,进一步凸显了多模态融合的必要性。

1.3 多模态分类

音乐情感表达具有复杂性和多维性,一首乐曲的情感不仅体现在音频时频信息和歌词语义之中,还可能涉及其他模态。多模态信息的融合利用,为提升音乐情感识别性能提供了新的研究路径。

当前研究中,歌词与音频的结合最为普遍。陶凯云[22]对传统 LFSM(Late Fusion Sub-task Merging)融合策略进行改进,通过音频与歌词融合实现多模态音乐情感分类,准确率提升至 84.43%,验证了多模态融合的有效性。Gao 等学者分别采用 CNN 和 BERT 预训练模型处理音频与歌词,并针对音频提取多种特征组合以确定最佳融合形式。实验表明,不同模态信息的协同利用可显著提升情感识别性能,而特征组合的选择直接影响最终效果。

多模态融合主要分为数据级、特征级和决策级三种形式。其中,决策级融合层级最高,但传统线性加权方式下,单一模态内不同类别被迫共用权重,难以针对不同情感类别动态分配权重,限制了融合效果。陈坤等学者借助神经网络改进决策级融合机制,性能较原有线性加权提升 6.4%,为多模态融合提供了新思路,但提升幅度仍有限。

早期多模态研究、如 Morency 等学者[23]从视频中分别提取图像与音频信息进行分段融合,虽未取得理想效果,却为后续工作提供重要启示。Simpson 等学者[24]尝试将音频信号图转化为频谱图并借助图像识别技术处理,但本质仍属单模态范畴,未能实现真正意义上的多模态融合。近年来,深度融合的研究日益增多。陈炜亮[25]提出二次融合的多模态深度学习情感分析框架,从音频提取 MFCC 特征、从歌词提取 MIDI 文本特征,经二次融合生成情感表征向量,实现了更精准的情感判别。

当前研究热点还包括深度学习与其他技术的结合,情感识别准确率有望进一步提升。然而,现有多模态方法多依赖简单特征拼接、线性加权或时序建模,难以有效处理音频(时频连续)与歌词(语义离散)之间的异步性和异构性,长程互补信息捕捉不足,特征交互仍不充分。这些局限性为本文提出基于异构图注意力网络的 SongGATNet 模型提供了重要研究依据。

2 方法

本研究提出了一种新的音乐情感特征识别模型—SongGATNet,图 1 显示了本文提出的音频和文本特征融合模型的整体架构。该模型以音频与歌词两类异构模态为输入,通过“单模态表征学习-异构图建模-图注意力融合-判别优化”四个阶段,实现对音乐情感的精细化建模与分类。

具体而言,首先针对音频模态,采用预训练的 Audio Spectrogram Transformer(AST)模型提取时频特征;针对歌词模态,则使用 BERT+CNN 混合模型提取语义与局部情感特征。这 2 个单模态特征提取模块为后续融合提供了高质量的异构输入。其次,在特征融合阶段,将音频片段与歌词片段显式建模为异构图结构,通过跨模态全连接和同模态全连接构建细粒度交互关系,并采用双层 GAT 编码器实现注意力驱动的跨模态信息融合,同时引入混合全局池化(均值池化与最大池化)生成稳健的歌曲级表征。最后,通过一个强化分类头(包含 Batch Normalization、ELU 激活和 Dropout 机制)对融合后

的歌曲级特征进行四分类，并结合加权损失函数进一步提升模型对类别不平衡的适应能力。

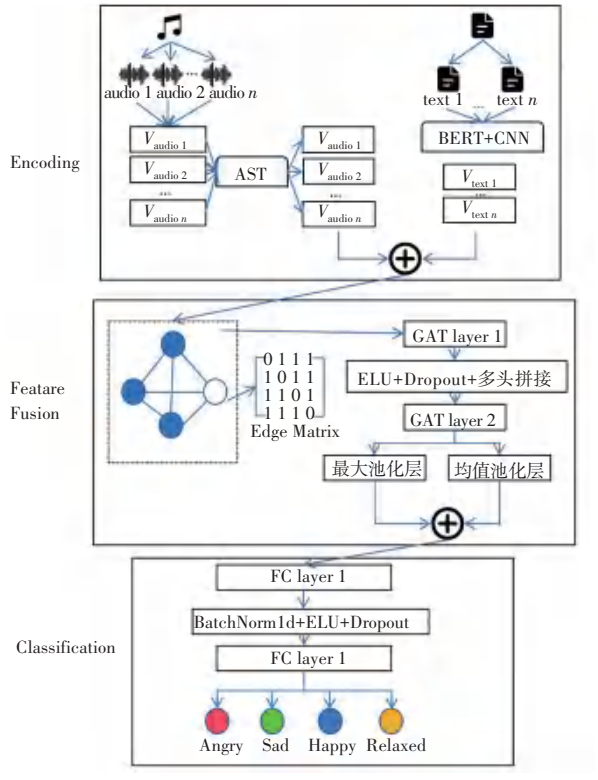

图 1　用于多模态音乐情感分类的 SongGATNet 总体架构

Fig. 1　Overall architecture of SongGATNet for multimodal music emotion classification

上述 4 个模块相互衔接、层层递进，共同构成了 SongGATNet 从单模态特征提取到多模态深度融合再到最终分类的完整技术流程。以下将分别对各模块进行详细阐述。

2.1　音频特征提取

音频模态包含丰富的时频信息，是音乐情感表达的重要载体。本文采用预训练的 Audio Spectrogram Transformer（AST）模型进行音频特征提取。AST 由 Gong 等学者[26]提出，是首个纯注意力机制（无卷积层）的音频分类模型，具体通过将音频信号转换为二维 log-Mel 频谱图后，直接应用 Vision Transformer 架构进行处理，能够有效捕捉长程全局时频依赖关系。

具体流程如下：首先，将每首歌曲切分为固定长度为 10 s 的音频片段；然后，对每个片段计算 log-Mel 频谱图作为输入。log-Mel 频谱图的计算公式为：

$$S_{\log-\mathrm{mel}}(t,m)=\log\left(\sum_{k}\left|\mathrm{STFT}\{x(t)\}(k)\right|^{2}\cdot H_{m}(k)\right) \tag{1}$$

其中，$x(t)$ 表示原始音频信号；$\mathrm{STFT}\{x(t)\}(k)$ 表示第 t 帧在频率 bin$k$ 上的短时傅里叶变换结果；$H_m(k)$ 表示第 m 个 Mel 滤波器的权重；m 表示 Mel 频率 bin 的索引；对数运算 log 用于压缩动态范围，使其更符合人耳的感知特性。

然后，将 log-Mel 频谱图输入预训练 AST 模型，获得每个音频片段的高维特征向量（维度 768）。在单模态验证阶段，附加一个简单的 MLP 分类头进行四分类（Sad、Angry、Happy、Relaxed），以评估特征质量。其分类过程可表示为：

$$\hat{y}=\mathrm{Softmax}(\boldsymbol{W}_2\cdot\mathrm{ReLU}(\boldsymbol{W}_1\cdot\boldsymbol{h}+\boldsymbol{b}_1)+\boldsymbol{b}_2) \tag{2}$$

其中，$\boldsymbol{h}$ 表示 AST 输出的全局特征向量；$\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 分别表示第一层全连接层的权重和偏置；$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 分别表示第二层全连接层的权重和偏置；ReLU 表示激活函数，用于引入非线性；Softmax 用于将输出转换为 4 类情感的概率分布。

为增强模型对噪声和变异的鲁棒性，训练过程中引入多种数据增强技术，包括添加高斯噪声、随机音量扰动、时间拉伸以及音高移位。同时，采用 AdamW 优化器并设置早停机制，避免过拟合。

通过预训练 AST 模型提取的音频特征不仅保留了丰富的时频谱信息，还能有效捕捉长程依赖关系，为后续 SongGATNet 模型中异构图的音频节点特征初始化提供了高质量输入。

2.2　歌词特征提取

歌词作为音乐作品中重要的语义载体，蕴含丰富的情感词汇和上下文信息，能够有效弥补音频模态在语义理解上的不足。本文采用 BERT+CNN 混合模型进行歌词特征提取。其中，BERT（Bidirectional Encoder Representations from Transformers）负责提取全局语义表征，而多核 CNN 则用于捕捉局部情感特征。

具体实现流程如下：首先，对原始歌词文本进行分词、去除停用词等预处理操作；然后，将处理后的歌词序列输入预训练的 BERT 模型，获得每个 token 的上下文相关嵌入向量；接着，在 BERT 输出的基础上叠加多核 CNN 模块，卷积核大小分别设置为 3、4、5，以提取不同尺度的局部情感模式；最后，在歌曲级预测阶段采用软投票机制，将多个歌词片段的预测结果聚合为整首歌曲的情感标签。

该模型的局部特征提取过程可表示为：

$$\boldsymbol{c}_i^k=\mathrm{ReLU}(\boldsymbol{W}^k\cdot\boldsymbol{h}_{i:i+k-1}+\boldsymbol{b}^k),k\in\{3,4,5\} \tag{3}$$

$$\boldsymbol{f}^k=\max_i(\boldsymbol{c}_i^k) \tag{4}$$

其中，$\boldsymbol{h}_{i:i+k-1}$ 表示 BERT 输出的连续 k 个 token 的隐状态序列；$\boldsymbol{W}^k$ 和 $\boldsymbol{b}^k$ 分别表示第 k 个卷积核的权重矩阵和偏置；$\boldsymbol{c}_i^k$ 表示第 k 个核在位置 i 处的卷

积特征；f^k 表示最大池化后得到的第 k 个核的显著局部特征向量。最后，将不同核尺寸提取的特征 f^3、f^4、f^5 拼接，再与 BERT 的全局表征结合，输入分类层。通过 BERT+CNN 模型提取的歌词特征为 SongGATNet 的异构图节点提供了丰富的语义信息，与 AST 提取的音频时频特征形成良好互补，为后续跨模态注意力融合奠定了坚实基础。

2.3 基于异构图注意力的特征融合

传统多模态音乐情感识别方法多采用简单特征拼接或线性加权，难以有效捕捉音频与歌词之间的异步性和异构性，导致跨模态互补信息利用不充分。为解决这一问题，本文提出 SongGATNet 模型，通过异构图注意力网络实现音频与歌词特征的深度融合。

SongGATNet 的核心思想是将单模态提取的特征显式建模为异构图结构，实现细粒度跨模态交互。具体流程如下：

首先，基于提取的特征，对于每首歌曲构建一个无向异构图 $G=(V,E)$。节点集合 V 由 2 部分组成：音频片段节点（数量为 N_a）和歌词片段节点（数量为 N_l）。每个节点的初始特征直接来自预训练 AST 模型提取的音频特征或 BERT+CNN 模型提取的歌词特征，特征维度均为 768 维。边集合 E 包含 3 类全连接边：跨模态边（每个音频片段节点与每个歌词片段节点之间建立双向连接，实现音频与歌词的直接信息交互）、音频同模态边（音频片段节点之间相互连接）和歌词同模态边（歌词片段节点之间相互连接）。这种异构图结构显式建模了模态内和模态间的关联关系，避免了传统拼接方法中特征维度爆炸和信息丢失的问题。其图构建可形式化表示为：

$$V=\{v_1^a,\cdots,v_{N_a}^a\}\cup\{v_1^l,\cdots,v_{N_l}^l\} \tag{5}$$

$$E=E_{\text{cross}}\cup E_{\text{audio}}\cup E_{\text{lyric}} \tag{6}$$

其中，$v_{N_a}^a$ 表示第 N_a 个音频节点；$v_{N_l}^l$ 表示第 N_l 个歌词节点；E_{cross} 表示跨模态边集合；E_{audio} 表示音频同模态边；E_{lyric} 表示歌词同模态边。

在构建好的异构图上，SongGATNet 采用双层 GAT 作为编码器进行特征融合。第一层 GATConv 使用多头注意力（heads=8），第二层使用单头注意力，通过残差连接和 ELU 激活增强特征传播能力。GAT 的注意力机制能够动态计算邻居节点的重要性权重，实现自适应跨模态信息聚合。其注意力系数计算公式为：

$$\alpha_{ij}=\frac{\exp(\text{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\parallel\boldsymbol{W}\boldsymbol{h}_j]))}{\sum_{k\in N(i)}\exp(\text{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\parallel\boldsymbol{W}\boldsymbol{h}_k]))} \tag{7}$$

其中，$\boldsymbol{h}_i$，$\boldsymbol{h}_j$ 表示节点 i 和节点 j 的初始特征向量；$\boldsymbol{W}$ 表示可学习的线性变换矩阵；$\boldsymbol{a}$ 表示注意力向量。

经过双层 GAT 编码后，对节点特征进行混合全局池化：同时采用全局均值池化和全局最大池化，并将两者拼接得到歌曲级表征。

2.4 分类与优化策略

经过双层 GAT 编码器和混合全局池化后，模型得到固定维度的歌曲级表征。为了进一步提升分类性能并增强泛化能力，本文设计了一个强化分类头。具体结构如下：首先将混合池化后的特征向量（维度为 hidden_channels×2，即 256 维）输入一个全连接层，映射到 256 维中间特征；然后依次接入 Batch Normalization、ELU 激活函数和 Dropout；最后通过一个线性层输出 4 维分类 logit。

该分类头采用强化设计主要考虑如下 3 点：

（1）Batch Normalization 的引入能够加速训练收敛并稳定梯度分布，有效缓解内部协变量偏移。

（2）ELU 激活函数相比 ReLU 能更好地处理负值区域，避免“死亡神经元”问题，同时保持梯度平滑。

（3）较高 Dropout 率（0.6）结合权重衰减，有效防止模型在有限训练样本（500 首歌曲）上的过拟合，提高泛化能力。

最终，模型使用加权交叉熵损失函数进行端到端训练，对 Happy 和 Relaxed 类别赋予较高权重，以缓解数据集中的类别不平衡问题。分类头输出经过 Softmax 后得到各类别概率。

3 实验

3.1 实验设置

3.1.1 数据集和评估指标

本实验采用 NJU-MusicMood 公开数据集[27]，该数据集共包含约 765 首英文歌曲，涵盖 Sad、Angry、Happy、Relaxed 四类情感标签。研究人员专门寻找被用户标注标签次数最多的音乐代表歌曲，有 Take This Heart Of Mine、Shelter 等知名歌曲，歌曲时长主要分布在 3~4 min。

数据集按歌曲级进行划分，确保训练集与测试集无重叠，从而避免数据泄露。具体划分如下：训练集 500 首歌曲，测试集 265 首歌曲。音频特征由预

训练 AST 模型提取，歌词特征由 BERT+CNN 模型提取，每首歌曲对应多个 10 s 音频片段和歌词片段。

评估指标方面，使用了准确率、宏平均精确率、召回率、*F*1 分数作为评估指标。

3.1.2　实验参数设置

优化器采用 Adam，初始学习率为 0.000 05，权重衰减系数设置为 0.01。损失函数使用 Focal Loss，以更好地处理类别不平衡问题。训练过程中引入 *L*2 正则化和梯度裁剪，以稳定训练过程并防止梯度爆炸。学习率通过 ReduceLROnPlateau 调度器动态调整，监控验证集宏平均 *F*1 分数，衰减因子为 0.5，耐心值 patience 设为 15。批量大小设置为 8，最大训练轮数为 100 轮。实验采用固定随机种子 42，以确保结果可重复性。

为缓解数据集中的类别不平衡问题，训练时对 Happy 和 Relaxed 类别设置较高损失权重 1.3，其余类别权重为 1.0。同时，使用 PyTorch Geometric 的 DataLoader 构建图数据批次，训练集批次大小为 8，测试集批次大小为 1。模型训练在支持 CUDA 的 GPU 环境下进行，采用 PyTorch 和 PyTorch Geometric 框架实现。所有实验均在训练结束后，于测试集（265 首歌曲）上以评估模式进行最终推理和指标计算。

3.2　实验结果和讨论

3.2.1　实验结果

为全面验证所提 SongGATNet 异构图注意力多模态融合模型的有效性与优越性，本研究设计并实现了 3 组具有代表性的对比实验，分别代表当前主流的多模态融合策略。

（1）简单特征拼接融合（Concat+MLP）。最经典的早期多模态融合方式。将音频特征和歌词特征直接在特征维度上拼接得到 1 536 维向量，再输入 3 层 MLP，隐藏层维度从 1 024 维度到 512 维度再到 256 维度，最后接 Softmax 输出四分类结果。该方法不包含任何显式的跨模态交互机制，是最强的“无交互”基线。

（2）CNN-LSTM 融合模型。经典的时序建模+局部特征提取组合架构。音频采用 3 层 2DCNN 处理 log-Mel 频谱图得到序列特征；歌词使用 BERT 获得 token 序列表征；2 种模态特征在序列长度对齐后沿特征维拼接，输入双层双向 LSTM（隐藏单元 256），最后经全局平均池化+全连接层完成分类。该模型在早期多模态音乐情感识别文献中被广泛采用，具有较强的代表性。

（3）SongGATNet 的边稀疏化变体，通过计算所有节点对（跨模态+同模态）的余弦相似度，仅保留相似度大于 0.55 的边，并限制每个节点最多连接 12 个邻居。该变体边数量平均减少约 65%，能够大幅减少训练时间。

所有对比实验均严格采用与 SongGATNet 相同的配置。不同模型在测试集上的性能对比结果见表 1。不同模型在测试集上的性能对比结果如图 2 所示，每个类所对应模型预测准确的个数如图 3 所示。

表 1　不同模型在测试集上的性能对比

Table 1　Performance comparison of different models on the test set

标签	SongGATNet	Sparse-GAT	CNN-LSTM	Concat+MLP
Sad	0.93	0.90	0.86	0.85
Angry	0.94	0.90	0.95	0.95
Happy	0.94	0.91	0.91	0.89
Relaxed	0.93	0.91	0.81	0.83
Accuracy	0.94	0.91	0.88	0.88
MacroAvg *F*1	0.94	0.91	0.88	0.88

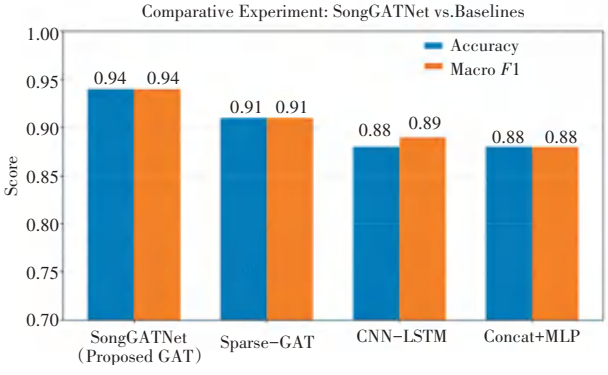

图 2　不同模型在测试集上的性能对比图

Fig. 2　Performance comparison of different models on the test set

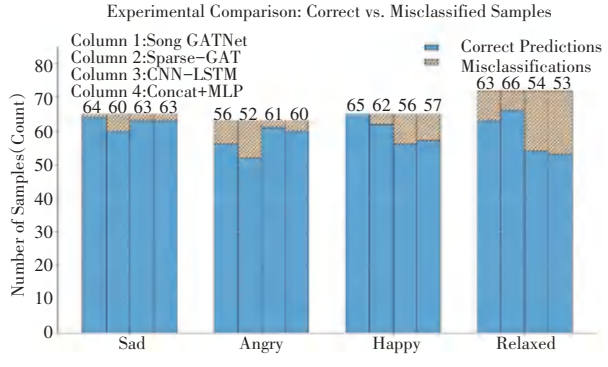

图 3　每个类所对应模型预测准确的个数

Fig. 3　Number of correctly predicted instances for each class

从图 2 可以看出，本文提出的 SongGATNet 在整体性能上取得了显著优势，准确率和宏平均 *F*1 分数均达到 0.94，较 3 种基线模型有明显提升。相比

简单拼接基线(Concat+MLP),SongGATNet 的 Macro $F1$ 从 0.88 提升至 0.94,提高了约 6%。这表明传统特征拼接方式无法有效利用音频与歌词之间的互补信息,而 SongGATNet 通过异构图结构实现了细粒度的跨模态交互,显著增强了特征表达能力。相比时序融合模型(CNN-LSTM),尽管 CNN-LSTM 在 Angry 类别上表现较强($F1 = 0.95$),但在 Relaxed 类别上明显薄弱($F1 = 0.81$)。SongGATNet 在 4 个类别上均取得了更为均衡且优异的性能($F1$ 均大于等于 0.93),说明双层 GAT 注意力机制比单纯的时序建模更能捕捉复杂的情感依赖关系。相比稀疏化变体(Sparse-GAT),Sparse-GAT 通过减少边数量在一定程度上降低了计算复杂度,但性能也随之下降(Macro $F1$ 从 0.94 降至 0.91)。这充分证明了全连接异构图(包含完整跨模态边和同模态边)对于充分挖掘音频歌词互补信息的重要性。

根据图 3 显示,SongGATNet 在 4 个情感类别上表现最为均衡,该模型在 Happy 类别达到了 100% 的识别率,并且在 Sad 类别上仅有 1 例误诊。相比之下,Concat+MLP 和 CNN-LSTM 在处理 Relaxed 和 Angry 情感时出现了明显的局限性,正确识别数在 53~60 之间波动。值得注意的是,SongGATNet 模型通过建模音频和歌词的全局拓扑关系,增强了模型对相似情感边界的区分能力,例如测试歌曲 Shelter,在前期的纯音频实验中,这首歌有着极高的平滑度和较少的动态起伏,往往被分类为 Relaxed 类,但是在歌词中存在着"what is getting you down",能够动态修正音频节点的误判,正确分类为 Sad 类。

总体而言,对比实验结果充分验证了 SongGATNet 的核心创新点—异构图建模 + 双层 GAT 注意力机制的有效性。该模型不仅在整体性能上超越了传统多模态融合方法,还在各类别均衡性和鲁棒性上展现出明显优势,为音乐情感识别任务提供了更强有力的技术支撑。

3.2.2 消融实验

为系统性地量化 SongGATNet 模型中各核心组件对音乐情感识别性能的贡献,本研究开展了全面的消融实验。所有消融实验均在与完整模型完全相同的实验设置下进行。消融设置具体如下。

(1)完整模型(基准):异构图结构+双层 GAT+混合全局池化+强化分类头。

(2)去除异构图(HomogeneousGraph):将所有音频片段节点与歌词片段节点统一视为同一种类型节点,不再区分跨模态/同模态边类型,图结构退化为普通同构图,其余组件保持不变。

(3)去除跨模态边(NoCross-ModalEdges):仅保留同模态内部边,完全切断音频与歌词之间的直接交互,其余结构不变。

(4)单层 GAT:仅保留第一层 GAT,移除第二层全局注意力聚合层,直接对第一层输出进行混合池化。消融实验结果见表 2,实验结果如图 4 所示。

表 2 消融实验结果

Table 2 Ablation experiment results

标签	SongGATNet	Ablation-Hom	Ablation-NoCM	Ablation-SL
Sad	0.93	0.93	0.87	0.68
Angry	0.94	0.97	0.96	0.94
Happy	0.94	0.92	0.90	0.85
Relaxed	0.93	0.92	0.83	0.81
Accuracy	0.94	0.93	0.89	0.83
MacroAvg $F1$	0.94	0.93	0.89	0.82

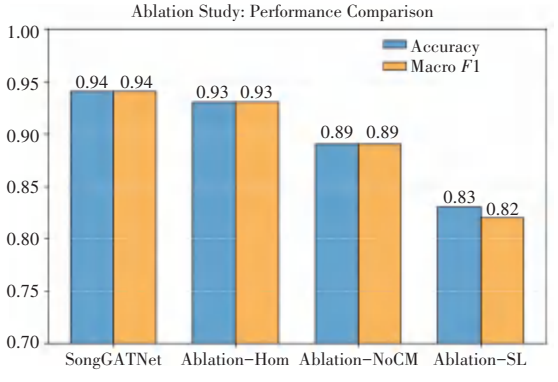

图 4 消融实验结果图

Fig. 4 Ablation study results

从实验结果可以清晰看出,不同模块对模型性能的影响程度存在显著差异:

首先,就是跨模态边的作用最为关键(Ablation-NoCM),当去除所有跨模态边后,模型性能下降最为明显,准确率从 0.94 降至 0.89,宏平均 $F1$ 从 0.94 降至 0.89,下降幅度达 0.05。这充分证明了音频与歌词之间的直接跨模态交互是 SongGATNet 性能提升的核心机制。没有跨模态边,模型难以有效融合 2 种模态的互补信息,导致在 Relaxed 和 Sad 类别上表现显著退化。

其次,就是双层 GAT 结构的必要性(Ablation-SL),将双层 GAT 简化为单层后,模型性能大幅下降,准确率降至 0.83,宏平均 $F1$ 降至 0.82,下降幅度达 0.12。其中,Sad 类别 $F1$ 从 0.93 骤降至 0.68,说明单层 GAT 无法充分捕捉长程全局依赖关系。双层结构通过逐层注意力传播,能够更好地实现细粒度特征交互和全局信息聚合,这是模型保持高性

能的重要保障。

最后，当去除同模态边时，性能仅轻微下降，Macro $F1$ 从 0.94 降至 0.93。这表明同模态边主要起到模态内信息平滑的作用，而跨模态交互才是主导因素。即使在同构图设置下，模型仍能维持较高性能，进一步验证了异构图中跨模态连接的核心价值。

综合消融实验结果可知，异构图中的跨模态全连接边和双层 GAT 注意力机制是 SongGATNet 取得优异性能的关键组件。混合全局池化与强化分类头则进一步提升了模型的稳定性和泛化能力。各模块相互配合，共同实现了音频时频特征与歌词语义特征的高效深度融合。

4　结束语

本文围绕音乐情感识别任务中多模态信息融合不足的问题，提出了一种基于图注意力网络的音频-歌词多模态情感分类模型 SongGATNet。分别利用 Audio Spectrogram Transformer（AST）与 BERT + CNN 模型获取音频时频表征与歌词语义表征，并通过构建音频片段与歌词片段的异构图结构，引入图注意力机制实现跨模态信息的动态交互与深度融合。相较于传统特征拼接或线性融合方法，所提出模型能够更有效地刻画模态间的异构性与非对齐关系，从而提升情感表示能力。

实验结果表明，在多种对比模型（包括 Sparse-GAT、CNN - LSTM 及 Concat + MLP 等）下，SongGATNet 在分类性能与模型稳定性方面均表现出明显优势。进一步的消融实验验证了异构图建模与图注意力融合机制在提升模型性能中的关键作用，表明该方法在跨模态特征交互建模方面具有良好的有效性与可扩展性。本文通过引入预训练模型进行特征提取，有效降低了模型训练成本与数据依赖程度，在中小规模数据场景下具备较强的实用价值。这一设计为多模态音乐情感识别任务提供了一种兼顾性能与效率的可行路径。

在未来研究中，可以尝试对预训练网络进行微调，或引入生理信号（如脑电、心率）等多模态数据，以进一步提升多模态音乐情感识别的精度，并探索其在中医情志调理、智能音乐推荐等实际场景中的应用价值。

参考文献

[1] 赵健谕. 音乐情感识别方法的研究[D]. 沈阳：辽宁大学，2011.

[2] YANG Y H, CHEN H H. Music emotion recognition[J]. Boca Raton：CRC Press，2011.

[3] 王军辉. 基于深度学习的音乐情感识别方法研究[D]. 大连：大连理工大学，2024.

[4] 康健，王海龙，苏贵斌，等. 音乐情感识别研究综述[J]. 计算机工程与应用，2022，58(4)：64-72.

[5] LARTILLOT O, TOIVIAINEN P, EEROLA T. A matlab toolbox for music information retrieval[C]// Proceedings of the 31st Annual Conference on Data Analysis, Machine Learning and Applications. Cham：Springer，2008：261-268.

[6] 纪正飚，王吉林，赵力. 基于模糊 K 近邻的语音情感识别[J]. 微电子学与计算机，2015，32(3)：59-62.

[7] 曹智贤. 基于机器学习的音乐情感识别方法研究[D]. 长沙：湖南大学，2018.

[8] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26(1)：36-46.

[9] 魏华珍，赵姝，陈洁，等. 特征组合的中文音乐情感识别研究[J]. 安徽大学学报(自然科学版)，2014，38(6)：30-36.

[10] 王秀，谢志成，张栋. 一种基于特征差异度和 SVM 投票机制的数字音乐语音情感识别算法[J]. 福州大学学报(自然科学版)，2015，43(4)：460-465.

[11] DOWNIE S J. The music information retrieval evaluation exchange (2005 - 2007)：A window into music information retrieval research[J]. Acoustical Science and Technology，2008，29(4)：247-255.

[12] YANG D, LEE W S. Disambiguating music emotion using software agents[C]//Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004). Barcelona, Spain：UPF，2004，4：218-223.

[13] SIGTIA S, DIXON S. Improved music feature learning with deep neural networks[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2014：6959-6963.

[14] CHOI K, FAZEKAS G, SANDLER M. Automatic tagging using deep convolutional neural networks[J]. arXiv preprint arXiv，1606.00298，2016.

[15] 陈长风. 基于 CNN-LSTM 的歌曲音频情感分类[J]. 通信技术，2019，52(5)：1114-1118.

[16] TANG H, CHEN N. Combining CNN and broad learning for music classification[J]. IEICE Transactions on Information and Systems，2020，E103-D(3)：695-701.

[17] THAMMASAN N, FUKUI K, NUMAO M. Application of deep belief networks in eeg-based dynamic music-emotion recognition[C]//Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN). Piscataway，NJ：IEEE，2016：881-888.

[18] CHEN R H, XU Z L, ZHANG Z X, et al. Content-based music emotion analysis and recognition[C]//Proceedings of 2006 International Workshop on Computer Music and Audio Technology. DMAC，2006：68-75.

[19] XIA Y, WANG L, WONG K F. Sentiment vector space model for lyric-based song sentiment classification[J]. International Journal of Computer Processing of Languages，2008，21(4)：309-330.

[20] YANG D, LEE W S. Music emotion identification from lyrics[C]//Proceedings of 2009 11th IEEE International Symposium on Multimedia. Piscataway，NJ：IEEE，2009：624-629.

[21] 杜常辉. 基于深度学习的歌词文本配图[D]. 哈尔滨：哈尔滨

工业大学,2020.
[22]陶凯云. 基于音频和歌词的音乐情感分类研究[D]. 南京:南京邮电大学,2015.
[23]MORENCY L P, MIHALCEA R, DOSHI P. Towards multimodal sentiment analysis: Harvesting opinions from the Web [C]//Proceedings of the 13th International Conference on Multimodal Interfaces. New York:ACM,2011: 169-176.
[24]SIMPSON A J R, ROMA G, PLUMBLEY M D. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network [C]// Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation. Cham: Springer, 2015: 429-436.
[25]陈炜亮. 音频文本混合的歌曲深度情感识别[D]. 合肥: 合肥工业大学, 2017.
[26]GONG Y, CHUNG Y A, GLASS J. Ast: Audio spectrogram transformer[J]. arXiv preprint arXiv, 2104.01778, 2021.
[27]XUE H, XUE L, SU F. Multimodal music mood classification by fusion of audio and lyrics[C]//Proceedings of the International Conference on Multimedia Modeling. Cham: Springer, 2015: 26-37.