计算机科学辑要
Compendium of Computer Science
ISSN Print: 3079-5362
ISSN Online: 3079-5370
https://doi.org/10.63313/CS.8027

基于通道自协调注意力和时频掩蔽数据增强的音乐流派分类

温治芳

湖南邮电职业技术学院，湖南省 长沙市

发布日期：2026年10月20日

摘　要

音乐流派分类是音乐信息检索的重要组成部分，得益于卷积神经网络在计算机视觉领域取得突破性进展，研究人员将其应用于音乐流派分类。卷积神经网络捕获局部特征时，卷积核将所有通道视为一个整体进行特征提取。然而，各个通道所携带的特征信息并非同等重要。因此，本文提出了通道自协调注意力，自适应地校准特征图中不同通道对全局信息的贡献度。同时，为了解决模型过拟合问题，本文采用随机掩蔽频谱图时频域的方法进行数据增强。实验结果表明：本文提出的模型在GTZAN数据集上取得了90.8%准确率。

关键词

音乐流派分类；卷积神经网络；注意力机制；时频域掩蔽；数据增强

Channel self-coordinated attention and time-frequency masking data augmentation based music genre classification

Zhifang Wen

Hunan Post and Telecommunication College, Changsha 410004, Hunan, China

Abstract

Music genre classification (MGC) is an indispensable part of music information retrieval. Because convolutional neural network (CNN) has made a breakthrough in computer vision, researchers apply it to MGC. When the CNN captures local features, the convolution kernel treats all channels as a whole for feature extraction, however, the feature information of each channel is not equally important. Therefore, this paper proposes channel self-coordinated attention (CSCA) to adaptively calibrate the contribution of different channels to the global information in the feature map. At the same time, in order to solve the model overfitting, this paper adopts the method of randomly

文章引用: 温治芳. (2026). 基于通道自协调注意力和时频掩蔽数据增强的音乐流派分类. 计算机科学辑要, 1(3), 43 - 50. ISSN Print: 3079-5362, ISSN Online: 3079-5370.
DOI: 10.63313/CS.8027

masking the time-frequency domain of the spectrogram for data augmentation. Experimental results show that the accuracy of our model is 90.8% on the GTZAN dataset.

Keywords

Music Genre Classification; Convolutional Neural Network; Attention Mechanism; Time-Frequency Domain Masking

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



1. 引言

音乐流派分类是根据歌曲内容将其划分为不同音乐流派的自动化方法，对音乐信息检索、音乐推荐等相关研究和应用具有重要意义。传统分类方法依赖于手工设计声学特征，使研究范围变得局限。深度学习的端到端学习模式，在输入端提供原始数据，输出端生成最终结果，无需单独考虑特征设计过程，这为音乐流派分类提供了可行、便捷的新思路。

随着多媒体和通信技术的高速发展，数字音乐在生活中随处可见。然而，音乐资源体量也变得越来越大，依靠人工对海量资源进行检索变得极为困难。因此，音乐信息检索[1]成为一个具有挑战性的问题。音乐流派作为区分音乐的高级标签之一，如果能够自动根据音乐内容将音乐划分成不同流派，对音乐信息检索会起到重要作用。由此可见，一个准确、有效的音乐流派分类系统是不可或缺的[2, 3]，其主要由两个部分组成：1）特征提取 2）分类器模型。传统方法[4, 5]研究音乐流派分类时依赖手工制作的特征，要求研究人员具有音乐领域的专业知识。近年来，深度学习在自然语言处理[6]和计算机视觉[7]等领域取得了突破性进展。深度学习的优点之一在于：它提供了一种端到端学习方式，不需要额外考虑特征提取和分类器模型。音频的低级特征[8]，如短时傅里叶变换频谱图（STFT-spectrogram）和梅尔频谱图（Mel-spectrogram）等，与 RGB 图像相似。研究人员使用深度学习中的卷积神经网络（convolutional neural network，CNN）模拟识别图像的方式对频谱图进行分类。

在音乐流派分类方向，Zhang 等人[9]提出了基于卷积神经网络结合平均池化、最大池化和跳跃连接的方法并采用短时傅里叶变换频谱图作为输入特征，提高了音乐流派分类的识别精度。在此基础上，为了捕获音频的时序依赖性，Choi 等人[10]提出了卷积循环神经网络（CRNN）用于音乐分类，其中卷积神经网络提取局部特征，递归神经网络对提取的特征进行时序语义累积。深度学习研究方法大多都基于卷积神经网络，然而其通过卷积核提取局部特征时，卷积核独立地完成对特征图中所有通道的特征信息提取，使得卷积后特征图中通道之间产生信息隔离，然而并非所有通道的特征信息都同等重要。受挤压激励网络等工作[11, 12]的启发，本文提出了通道自协调注意力（channel self-coordinated attention，CSCA），其通过构造表征通道间关系的权重矩阵，对通道特征因子进行加权求和获得通道间的权重系数，自适应校准特征图中不同通道的贡献程度。通道自协调注意力通过全局池化函数将特征图压缩成通道因子向量，因此，本文比较了全局平均池化和全局最大池化两种池化函数的作用。权重矩阵中特征因子的非线性激活由激活函数来执行，本文讨论了 Sigmoid、ReLU 和 Softmax 三种激活函数。值得注意的是，通道自协调注意力为灵活替换模块、整体参数量小、不改变特征图形状，可嵌入网络的主干层。

根据 Kahl 等人[13, 14]的研究，本文采用轻量级卷积网络 CNN-5 识别音频转换而成的频谱图，并且

将通道自协调注意力嵌入网络主干层，构成基于通道自协调注意力的 CNN-5 网络，自适应地调整卷积后特征图中不同通道对全局信息的贡献程度。此外，基于深度学习的方法需要大量数据来训练模型，依靠误差反向传播算法不断调整模型中的参数，使得模型预测结果尽可能接近真实标签。因此，本文采用在线时频掩蔽数据增强[15]扩充数据集，在训练过程中不增加数据集规模，并且达到数据增强效果。主要贡献总结如下：

（1）为了突出携带重要特征信息的通道，本文提出了通道自协调注意力自适应地校准特征图中不同通道对全局信息的贡献度，并且讨论了全局池化函数和激活函数对注意力机制性能的影响。

（2）为了维持数据集规模恒定并且提升模型性能，本文采用了时频掩蔽数据增强即随机掩蔽频谱图的时频域特征。

（3）本文采用基于通道自协调注意力的 CNN-5 网络，并对数据进行时频掩蔽增强。在 GTZAN 数据集上的实验结果达到了 90.8%准确率。

2. 研究方法

如图 1 所示，模型总体结构分为三个部分，首先将歌曲经过短时傅里叶变换、时频域叠加、梅尔滤波器组和取对数等操作后，转换成梅尔频谱图。其次，将梅尔频谱图送入基于通道自协调注意力的 CNN-5 网络，提取频谱图中特征信息，全局平均池化层将特征图压缩成一列特征向量，最后特征向量经由全连接层，映射为十个类别的概率向量，选择预测概率最高的类别，当成输出结果。在本节中，主要介绍 CNN-5 网络、通道自协调注意力和时频掩蔽数据增强。

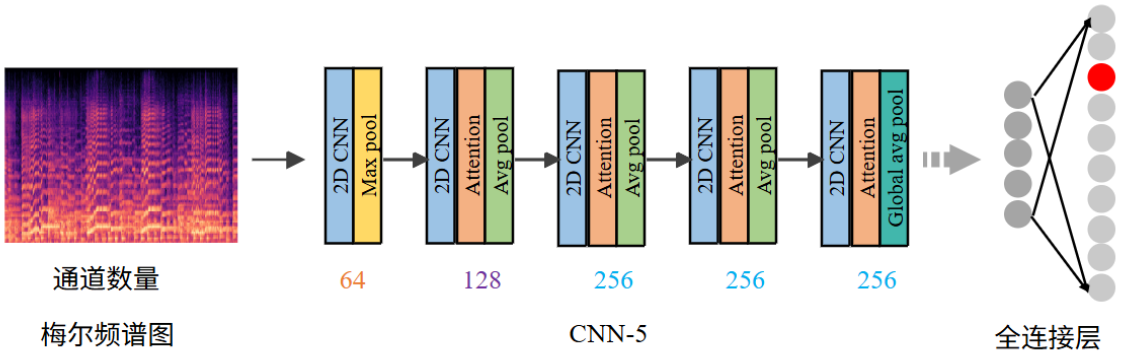


Figure 1. Overall architecture diagram
图 1. 总体结构图

2.1. CNN-5 网络

Kahl 等人发现，识别音频转换而成的频谱图时，经典结构的卷积网络比深层网络更有效，本节沿用此思路提出了 CNN-5 网络，其由五层卷积层构成。首先采用较少通道的卷积层捕获纹理和轮廓等低级特征，随着层数加深，增加通道数量累积深层语义信息。网络结构信息如下：

L1：第一层卷积层由 64 个卷积核构成，感受野大小为 7×7。首层设置大感受野来捕获更大范围内的频谱特征，接着采用最大池化完成下采样，保留池化块中显著特征信息。

L2：第二层卷积层由 128 个卷积核构成，感受野大小为 3×3。在第二层及之后的卷积层中下调感受野的大小，降低参数量。同时采用平均池化完成下采样，保留池化块中所有特征信息。

L3：第三层卷积层由 256 个卷积核构成，感受野大小为 3×3。采用平均池化做下采样。

L4：第四层卷积层由 256 个卷积核构成，感受野大小为 3×3。采用平均池化做下采样。

L5：第五层卷积层由 256 个卷积核构成，感受野大小为 3×3。采用全局平均池化做下采样，将单个通道中所有特征信息，通过计算平均值的方式压缩到一个单位中，避免信息冗余同时减少全连接层时的网络参数量。全局平均池化后特征图形状与通道数量一致。

Batch Normalization[16]设置在每个卷积层之后，加速网络的训练。网络采用 ReLU 函数[17]进行非线性激活。

2.2. 通道自协调注意力

为了消除通道间的信息隔离，自适应地调整不同通道对特征图的重要性。本文提出了通道自协调注意力（如图 2 所示），它通过构造通道信息关联矩阵表征任意两个通道之间的相互关系，并与通道因子向量进行矩阵乘法加权求和，获得表征通道关系的特征向量后，将特征向量扩张成与特征图形状一致，并与特征图逐元素相乘从而完成特征图中通道信息的差异性校准。

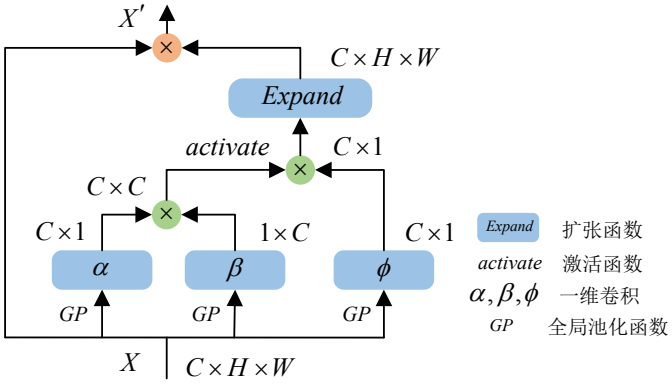


Figure 2. Channel self-coordination attention structure
图 2. 通道自协调注意力结构图

给定输入特征图 $X \in R^{C\times H\times W}$，其中 C 表示通道数量，H 表示特征图的高，W 表示特征图的宽。通道自协调注意力的具体过程表示如下：

$$F_{Att}(X) = F_{Expand}(\theta(\alpha(GP(X))\beta(GP(X)^{\mathrm{T}}))\phi(GP(X)))\cdot X \tag{1}$$

其中 GP 表示全局池化函数，θ 表示激活函数，F_{Expand} 表示扩张函数以及 α, β, ϕ 表示一维卷积。首先，采用 GP 将单个通道中所有特征信息压缩成一个通道因子。挑选全局平局池化函数介绍池化操作，仅选择一个通道 X_i 表示此过程，如下所示：

$$GP(X_i) = \frac{1}{H\times W}\sum_{p=1}^{H}\sum_{q=1}^{W}X_i(p,q) \tag{2}$$

$GP(X)$ 表示特征图中所有通道压缩后的通道因子向量并采用一维卷积调整通道因子间关系表示为 $\alpha(GP(X))$，获得另一个通道因子向量将其转置表示为 $GP(X)^{\mathrm{T}}$ 同样加上一维卷积，表示为 $\beta(GP(X)^{\mathrm{T}})$。将 $\alpha(GP(X))$ 和 $\beta(GP(X)^{\mathrm{T}})$ 进行矩阵乘法，得到表征任意两个通道间相互关系的通道信息关联矩阵。接着，对通道信息关联矩阵进行激活，将表征通道间关系的特征因子进行非线性激活后表示为 $\theta(\alpha(GP(X))\beta(GP(X)^{\mathrm{T}}))$。同样，将特征图中通道进行压缩，得到通道因子向量 $GP(X)$，加上一维卷积表示为 $\phi(GP(X))$。然后，将通道信息关联矩阵与通道因子向量进行矩阵乘法，使得 $\phi(GP(X))$ 中每一个通道特征因子与信息矩阵的每一行对应相乘并相加，使得所有通道特征因子相互关联。接着将 $\theta(\alpha(GP(X))\beta(GP(X)^{\mathrm{T}}))\phi(GP(X))$ 通过 F_{Exapnd} 扩张成与特征图尺寸一致。最后，将扩张后的特征因子与特

征图 X 逐通道相乘，校准不同通道对特征图的贡献程度。

2.3. 时频掩蔽数据增强

深度学习方法需要大量数据训练模型，有限的数据容易导致模型过拟合等问题。传统数据增强方法通过扩充数据集规模，制造更多数据来训练模型解决此问题。然而随着数据集规模的增大，模型训练时间也会增加。本节介绍的时频掩蔽数据增强，在训练过程中采用随机掩蔽输入特征，即随机掩蔽频谱图的时频域（如图 3 所示），为模型在训练时制造干扰元素，避免扩充数据集规模，同时又提升模型性能。假设梅尔频谱图中梅尔滤波器组的数量表示为 f，时序帧数量表示为 t，频率超参数表示为 ω，时序超参数表示为 τ。掩蔽时频域数据增强分为掩蔽频域，掩蔽时域和掩蔽时频域三种形式。掩蔽时频域为在掩蔽频域基础上，再掩蔽时域。因此仅介绍掩蔽频域和掩蔽时域，如下所示：

（1）掩蔽频域：假设 f_R 为连续掩蔽的梅尔频率，f_R 从 0 到 ω 的均匀分布中随机生成，表示为 $f_R \in (0,\omega)$。f_0 为频率掩蔽起始点，从 0 到 $f - f_R$ 中选择，表示为 $f_0 \in (0, f - f_R)$。掩蔽频域的范围是 $(f_0, f_0 + f_R)$。

（2）掩蔽时域：假设 t_R 为连续掩蔽的时序步长，t_R 从 0 到 τ 的均匀分布中随机生成，表示为 $t_R \in (0,\tau)$。t_0 为时序掩蔽起始点，从 0 到 $t - t_R$ 中选择，表示为 $t_0 \in (0, t - t_R)$。掩蔽时域的范围是 $(t_0, t_0 + t)$。

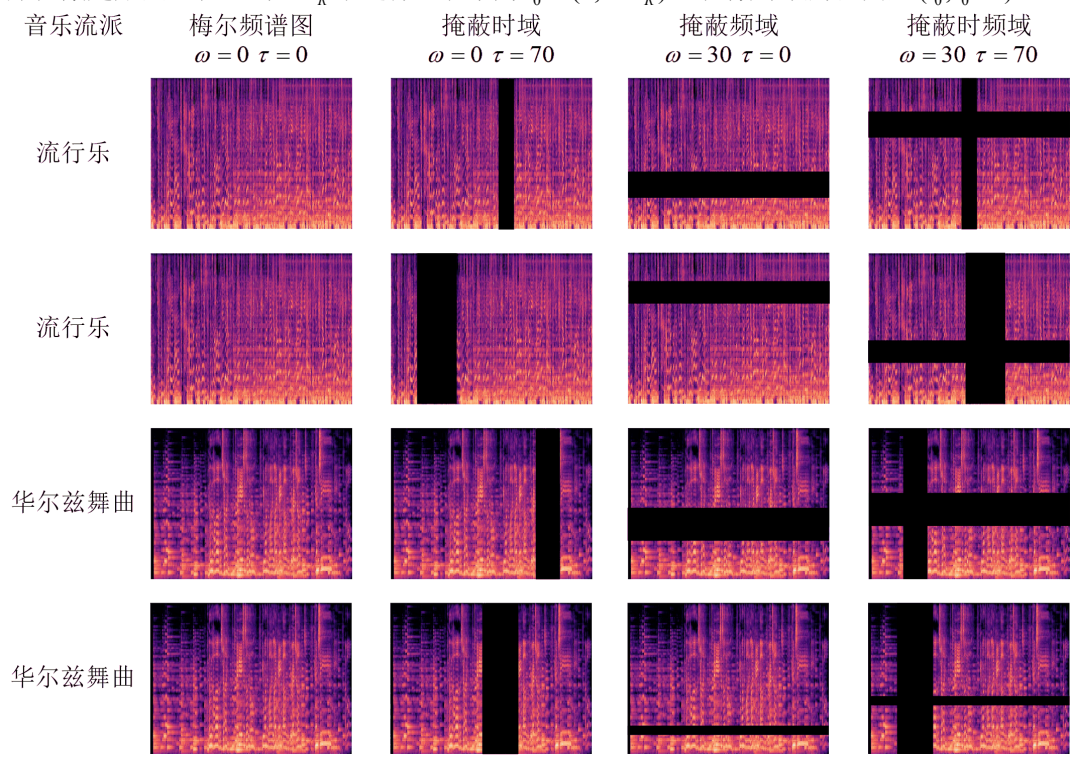


Figure 3. Time-frequency masking data augmentation
图 3. 时频掩蔽数据增强示意图

3. 实验结果与分析

3.1. 数据集和预处理

本文使用的数据集是由 G. Tzanetakis 收集的 GTZAN[18]，在音乐流派分类方向被广泛地应用，极具

代表性。它包含 1000 首歌曲，均匀地分布在十个流派中，每个流派 100 首歌曲，具体流派分别为 Blues、Classical、Country、Hiphop、Jazz、Metal、Pop、Reggae 和 Rock。每首歌曲节选大约为 30 秒，采样频率为 22050 Hz，16 位。本文将 GTZAN 中的歌曲，采样率统一调整为 16000 Hz，此操作将歌曲转成单声道处理，避免原始歌曲中因多声道造成的信息重复。

本文将歌曲转换成梅尔频谱图，令其为输入特征，其中短时傅里叶变换窗口长度为 512，窗口跳跃大小为 256，频率分箱数为 128。GTZAN 数据集中数据量偏小，因此采用数据扩充操作。具体是将 GTZAN 中每一首歌曲采用滑动窗口算法切分成 11 个音乐片段（约 5 秒），每个切片之间重叠 50%，切分后的形状为 313×128。在实验中将训练集，验证集和测试集的比例划分为 8：1：1，并保持流派间的平衡。此外，单次实验结果的波动比较明显，本文采用十折交叉验证确保结果的稳定性，之后每次结果都是十次运行后的平均值。

3.2. 实验设置

本文采用 PaddlePaddle 框架，显卡型号为 Tesla V100。优化器为 Adam[19]，Batch size 为 33，损失函数为交叉熵（Cross Entropy），每一折数据训练 90 个 epochs，十折交叉验证一共训练 900 个 epochs。初始学习率为 0.0001，学习率衰减方法为余弦衰减。在训练阶段，我们将所有音乐片段全部打乱进行训练。在验证与测试阶段则是将同一首歌曲的所有片段连续排列进行验证和测试。并且，采用投票机制将所有音乐片段的输出值相加，选择一个概率最大的流派作为这首歌曲的最终输出。

3.3. 实验结果

Table 1. Comparison of experimental results by different models
表 1. 不同模型实验结果比较

模型	特征	准确率
KCNN(k=5) + SVM[18]	Mel-spectrum etc.	83.9 %
nnet2[8]	STFT	87.4 %
Transform learning[19]	Convnet features	89.8 %
Hybrid model[20]	MFCC, SSD, etc.	90.0 %
CNN-RNN with GLR[21]	Mel-spectrogram	87.8 %
MhaNN-SVM[22]	Mel-spectrogram	88.4 %
CNN-5	Mel-spectrogram	88.9 %
CNN-5 + CSCA (本文)	Mel-spectrogram	90.8 %

本节中，将基于通道自协调注意力的 CNN-5 网络与多种方法现有方法进行了比较，结果汇总在表 1。无论是早期的 KCNN+支持向量机（SVM 方法、基于卷积网络重新设计网络的 nnet2 方法、基于混合特征和迁移学习的 Transform learning 方法、基于混合模型的 Hybrid model 方法、基于全局正则化 CNN-RNN with GLR 方法以及基于多头注意力机制的 MhaNN-SVM 方法分类精度都低于本文提出的模型。

如图 4 所示，该混淆矩阵图呈现了基于通道自协调注意力的 CNN-5 网络在 GTZAN 数据集上预测结果与真实结果的对比。对角线的数值越大，颜色越深表示该音乐流派类别的准确率越高。古典乐（Classical）、金属乐（Metal）和爵士乐（Jazz）等音乐流派分类精度较高，尤其是古典乐分类精度达到 100%。而摇滚乐（Rock）和流行乐（Pop）等流派分类精度较低。从音乐曲风和器乐与人声构成角度分析：古典乐和爵

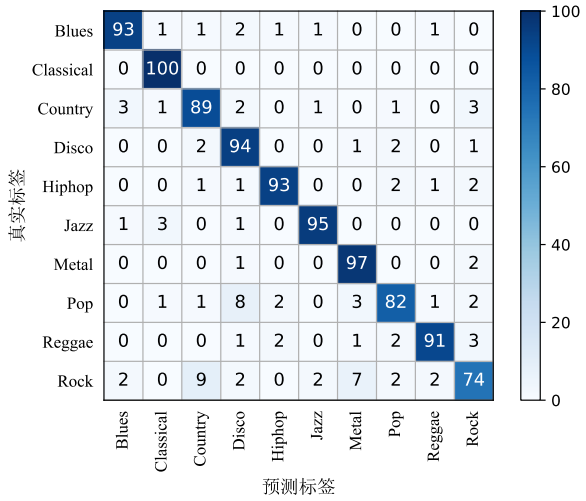


Figure 4. Confusion matrix of model
图 4. 模型混淆矩阵

士乐曲风规整，形式统一，同时人声占比较少。虽然金属乐内容丰富，但是曲风仍旧比较统一。因此都容易被模型所预测。值得注意的是：表 2 中古典乐的 Precision 指标不是最高，此指标是评价模型预测流派为古典乐的歌曲中，实际歌曲流派为古典乐的比重占多少。这表明：虽然古典乐容易被识别，但与其曲风类似的爵士乐（Jazz）同样容易被误分类为古典乐。摇滚乐和流行乐曲风多变人声占比突出，而人声不具有流派倾向。因此，导致了识别率低。正是由于摇滚乐和流行乐难以预测，其它流派歌曲被误分类为摇滚乐和流行乐的概率也很低，它们的 Precision 指标反而升高了。

Table 2. Analysis results of confusion matrix
表 2. 混淆矩阵分析结果

Classification	Precision	Recall	F1-score
Blues	0.94	0.93	0.93
Classical	0.94	1.00	0.97
Country	0.86	0.89	0.88
Disco	0.84	0.94	0.89
Hiphop	0.95	0.93	0.94
Jazz	0.96	0.95	0.95
Metal	0.89	0.97	0.93
Pop	0.9	0.82	0.86
Reggae	0.95	0.91	0.93
Rock	0.85	0.74	0.79

4. 结论

本文提出了基于通道自协调注意力的 CNN-5 网络用于音乐流派分类，CNN-5 网络用来提取频谱图的纹理和轮廓等低级特征以及高级语义信息，而通道自协调注意力通过自适应地校准卷积后特征图中不同通道的贡献程度，突出了携带重要特征信息的通道。本文还研究了全局池化函数和激活函数对通道自协调注意力的影响。全局平均池化函数通过平均池化操作保留了通道中所有的特征信息，效果优于全局最大池化函数。Sigmoid 函数单独对通道因子向量中每一个特征因子进行激活，激活后的数值范围恒定，在三种激活函数中，Sigmoid 函数性能最佳。此外，时频掩蔽数据增强也进一步提升了模型性能。本文提出

的模型属于监督学习方法需要人工标注数据，在未来的工作中无需标注数据的自监督学习值得关注。

基金项目

本研究获得湖南邮电职业技术学院校级科研课题《基于人工智能的电话客服语音情感识别研究》（课题编号：XJ24QY005）资助。

参考文献

[1] M. Schedl, E. Gómez, J. J. F. Urbano and T. i. I. Retrieval (2014). Music information retrieval: Recent developments and applications. 127-261

[2] E. Benetos, C. J. I. T. o. A. Kotropoulos, Speech, and L. Processing (2010). Non-negative tensor factorization applied to music genre classification. 1955-67

[3] S. Sigtia and S. Dixon (2014). Improved music feature learning with deep neural networks. IEEE, 6959-63

[4] B. K. Baniya, D. Ghimire and J. Lee (2014). A novel approach of automatic music genre classification based on timbrai texture and rhythmic content features. IEEE, 96-102

[5] A. F. Arabi and G. Lu (2009). Enhanced polyphonic music genre classification using high level features. IEEE, 101-06

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. J. A. i. n. i. p. s. Polosukhin (2017). Attention is all you need.

[7] A. Krizhevsky, I. Sutskever and G. E. J. A. i. n. i. p. s. Hinton (2012). Imagenet classification with deep convolutional neural networks.

[8] Z. Fu, G. Lu, K. M. Ting and D. J. I. t. o. m. Zhang (2010). A survey of audio-based music classification and annotation. 303-19

[9] W. Zhang, W. Lei, X. Xu and X. Xing (2016). Improved music genre classification with convolutional neural networks. 3304-08

[10] K. Choi, G. Fazekas, M. Sandler and K. Cho (2017). Convolutional recurrent neural networks for music classification. IEEE, 2392-96

[11] J. Hu, L. Shen and G. Sun (2018). Squeeze-and-excitation networks. 7132-41

[12] X. Wang, R. Girshick, A. Gupta and K. He (2018). Non-local neural networks. IEEE, 7794-803

[13] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter and M. J. C. Eibl (2017). Large-Scale Bird Sound Classification using Convolutional Neural Networks.

[14] Z. Wen, A. Chen, G. Zhou, J. Yi, W. J. M. t. Peng and applications (2024). Parallel attention of representation global time-frequency correlation for music genre classification. 83

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk and Q. V. J. a. p. a. Le (2019). Specaugment: A simple data augmentation method for automatic speech recognition.

[16] S. Ioffe and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. pmlr, 448-56

[17] V. Nair and G. E. Hinton (2010). Rectified linear units improve restricted boltzmann machines. 807-14

[18] G. Tzanetakis, P. J. I. T. o. s. Cook and a. processing (2002). Musical genre classification of audio signals. 293-302

[19] D. Kinga and J. B. Adam (2015). A method for stochastic optimization. California;, 1

[20] N. Karunakaran and A. Arya (2018). A scalable hybrid classifier for music genre classification using machine learning concepts and spark. IEEE, 128-35

[21] M. Ashraf, G. Geng, X. Wang, F. Ahmad and F. Abid (2020). A Globally Regularized Joint Neural Architecture for Music Classification. 220980-89 10.1109/access.2020.3043142

[22] L. Yang and H. Zhao (2021). Sound Classification Based on Multihead Attention and Support Vector Machine.