Computer Science and Application 计算机科学与应用, 2019, 9(10), 1892-1905
Published Online October 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.910212

Hans汉斯

Environmental Sound Classification Base on CNN and LightGBM

Weiping Liao, Pinghua Chen, Cong Zhao, Liang Zhao, Jianbing Chen, Mengqin Dong

Department of Computer, Guangdong University of Technology, Guangzhou Guangdong
Email: yepingweiyu@foxmail.com

Received: Sep. 25th, 2019; accepted: Oct. 10th, 2019; published: Oct. 17th, 2019

Abstract

Aiming at the problem that the traditional convolutional neural network has insufficient generalization ability and low accuracy in environmental sound classification, a new model mixing deep CNN with LightGBM is proposed. Based on the preprocessing of the Mel Frequency cepstral coefficient matrix on the audio file, the new model firstly uses the deep convolutional neural network to extract features. Then, combined with the efficient and accurate characters of LightGBM in classification prediction, the extracted features are imported into LightGBM for training. Thereby it achieves the purpose of improving classification accuracy. The results of the comparative experiments on the UrbanSound8K public dataset show that the new model improves the accuracy of 7.7% compared to the using a single-use convolutional neural network model.

Keywords

Environmental Sound Classification, Convolutional Neural Network, LightGBM Model, Mel Frequency Cepstrum Coefficient

基于CNN和LightGBM的环境声音分类

廖威平，陈平华，赵　璁，赵　亮，陈建兵，董梦琴

广东工业大学计算机学院，广东 广州
Email: yepingweiyu@foxmail.com

收稿日期：2019年9月25日；录用日期：2019年10月10日；发布日期：2019年10月17日

摘　要

针对传统卷积神经网络在环境声音分类中泛化能力不足且准确率不高的问题，提出了一个新的将CNN和

文章引用: 廖威平, 陈平华, 赵璁, 赵亮, 陈建兵, 董梦琴. 基于 CNN 和 LightGBM 的环境声音分类[J]. 计算机科学与应用, 2019, 9(10): 1892-1905. DOI: 10.12677/csa.2019.910212

LightGBM融合的环境声音分类模型。新模型在对音频文件进行梅尔频率倒谱系数矩阵预处理基础上，首先应用深度CNN提取音频的高层次特征；然后，结合LightGBM在分类预测上高效准确的特点，将提取的高层次特征导入LightGBM进行训练预测，从而达到提升分类准确性的目的。UrbanSound8K公开数据集上的对比实验结果表明：与目前使用的单独使用卷积神经网络相比，新模型提高了近7.7%的分类准确率。

关键词

环境声音分类，卷积神经网络，LightGBM模型，梅尔频率倒谱系数

1. 引言

音频识别是模式识别领域一个前沿研究课题。作为音频识别的一个重要分支，环境声音分类(Environmental Sound Classification, ESC)受到了业界许多专家学者的关注，成为了热门研究话题。ESC 是用机器进行声学环境分析最重要的技术之一，广泛应用于监听[1]、智能家居[2]、场景分析[3]和机器视听[4]等领域，如监管系统通过检测监管区域异常声音自动报告紧急情况并启动应急方案[5]、机器人通过对环境声音的分类识别确定下一步行动计划[6]等。与语音和音乐不同，环境声音的音频具有多样性特点，拥有更广泛的频率范围。近年来，随着医疗保健、安全监控、生态环境变化预测等应用需求的涌现，环境声音分类识别研究已越来越受到学术界的重视。环境声音的准确分类识别已成为相关应用成功与否的关键。

环境声音分类识别属于音频识别范围。传统的音频识别方法分为信号处理方法和机器学习方法。传统的信号处理方法直接使用音频数据[7] [8] [9]，如 Mel 滤波器组属性[10]、Gammatone 属性[11]、基于小波的属性[12]和多频带谱减法[13]等；传统的机器学习方法如 SVM [14] [15]、GMM [16]和 KNN [17]等。近年来，随着深度学习技术的发展，将深度神经网络(Deep Neural Network, DNN)应用于自动语音识别(Automatic Speech Recognition, ASR)和音乐信息检索(Music Information Retrieval, MIR) [18] [19]取得了巨大的成功。对于音频信号，DNN 能够从原始数据中提取特征，一些基于 DNN 的模型被提出并且表现得比传统的机器学习模型效果更好[20]，如：Picza K.J.将简单的卷积神经网络层结构应用于 log 梅尔频谱图，对环境声音进行分类处理[21]；Medhat F.等人通过嵌入类滤波器组的稀疏性来引导网络在频谱中的学习[22]；Takahashi 等人通过使用 log 梅尔频谱图和增量及增量的增量信息作为类似于图像 RGB 输入的三通道输入[23]。然而，DNN 的深度全连接架构对于转换特征并不具备强鲁棒性。一些新的研究发现卷积神经网络具有强大的通过大量训练数据探索潜在的关联信息能力，通过从环境声音中学习类似频谱图的特征[24]，将 CNN 应用于 ESC 的几次尝试已经获得了性能提升，如 Zhang 等人通过调整网络中各层的激活函数提高了模型的性能[25]；Zhang 等人通过调整卷积网络层结构并且融合混合样本生成新样本训练网络，提升了模型效果[26]。但是网络结构的设计依然有待改进，模型的特征获取与分类预测功能没有很好地进行分离，这为进一步改进模型提供了新的思路。

为了更好地利用音频数据信息，设计更好的网络结构模型，本文在此基础上调整网络层结构，同时由于卷积神经网络结构模型具有提取音频特征功能作用，本文将使用卷积神经网络模型对音频数据特征进行提取，而使用 LightGBM 模型对提取特征后的音频数据进行分类预测以加强模型效果，将模型提取

特征功能和分类预测功能分离，以改善模型效果。本文将卷积神经网络模型和 LightGBM 模型融合，融合了卷积神经网络提取特征的功能和 LightGBM 分类预测的能力，各分模型对应不同的功能，使模型结构具有更好的分类预测效果。

2. 技术细节

2.1. 音频数据预处理

音频数据存储着音频的信号，这种信号是一种一维的时域信号，由多段频谱信号按时间排列表示每个时间段的帧信息。通过直观上观察频谱的分布信息很难得出频率变化的规律，并且难以通过其它模型对该数据进行处理以进行分类预测，同时需要将每个音频数据进行规整化处理使其统一标准，因此需要将音频的频谱信息转换成其它易于理解的形式进行进一步处理。

人的听觉系统是一个特殊的非线性感知系统，它对不同频率的信号有不同的听觉敏感度，通常使用梅尔频率表示人耳对于频率的感受度。梅尔频率是一种基于人耳对等距的音高变化的感官判断而定的非线性频率刻度，它与频率的关系如公式(1)所示。

$$\mathrm{mel}(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

在音频特征的提取上，人类听觉系统能够做得非常好，它不仅能提取出语义信息，而且能提取出声源特征。如果在音频识别系统中能模拟人类听觉感知处理特点，就有可能提高音频的识别率。

梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)是通过梅尔频率和频率之间的关系计算得到的频率频谱特征，是一种在自动语音识别和声纹识别等音频识别中广泛使用的特征。梅尔频率倒谱系数考虑到了人类的听觉特征，通过对 spectrogram 声谱图进行描述的音频数据进行处理，对分帧后的音频频谱信息通过坐标表示出来，得到一个随时间变化的短时频谱图，即描述单帧信号中能量分布的情况。其峰值表示语音的主要频率成分，称为共振峰(formants)，携带声音的辨识属性，利用共振峰可以识别不同的声音。通过提取频谱的包络(Spectral Evnelope)和频谱细节，得到每帧频谱的相关信息。对于每帧频谱信息提取的一维频域特征，将每帧的结果沿另一个维度堆叠起来，得到类似于一幅图的二维信号形式，这样就可以像处理图像一样处理音频数据了。其具体特征提取过程如图 1 所示，先将语音预加重，使信号的频谱变得平坦，保持在低频到高频的整个频带中；再分为多个帧，每个帧对应于一个频谱；然后加窗，将每帧乘以汉明窗，以增加帧左段和右段的连续性；通过短时快速傅里叶变换(Fast Fourier Transformation, FFT)，计算频率与振幅的关系，使时域信号转换为频域上的能量分布来观察；因为频域信号有很多冗余信息，需要滤波器组对频域的能量幅值进行精简，每一个频段用一个值来表示。通过公式(1)将能量幅值转化为人耳对声音感知的梅尔频率，即 Mel 滤波操作，如图 2 的映射操作，并通过离散余弦变换将能量信号集种到低频部分。其输出结果值，即是能够描述音频数据的低阶特征 MFCC 值。梅尔频率倒谱系数将线性频谱映射到基于听觉感知的 Mel 非线性数据，即可进行倒谱分析。通过提取音频文件的 MFCC 物理特征，将音频文件信息转换为矩阵信息，为后续模型训练测试做准备。

2.2. CNN 提取特征数据

通过对频谱分析，得到音频声谱图对应的特征数据，且格式符合使用卷积神经网络(Convolution Natural Network, CNN)进行模型拟合过程，可以用 CNN 进行进一步的高层次特征提取。

卷积神经网络是一种前馈神经网络，由多层卷积层及池化层以及有限数量的全连接层以及 softmax 输出层构成的神经网络结构。卷积层是通过一系列卷积核表征像素点之间的空间分布，将一个范围内的

所有像素点进行加权求平均；卷积核用矩阵表示特征值，代表高层语义信息。卷积层的作用是过滤输入数据的特征，以提取输入所具有的特性空间，通过多个卷积核捕获不同视觉模式。池化层是通过对图像进行下采样将样本大小进行缩放或重构，为下一步更精细的特征做准备。常用的池化方法有最大池化和平均池化。全连接层对层间所有神经元节点进行权值连接，softmax 输出层对应模型的分类结果。

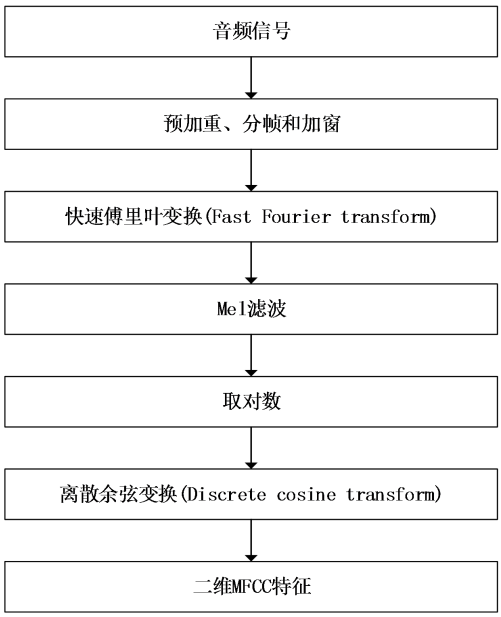

Figure 1. MFCC matrix extraction process for audio files
图 1. 音频文件的 MFCC 矩阵提取过程

卷积神经网络由在深层架构中堆叠在一起的许多不同层组成，包括一个输入层、一系列卷积层和池化层(可以任意方式组合)，防止过拟合的 Dropout 层，有限数量的完全连接的隐藏层，以及输出层。不同堆叠方式的网络结构，对模型有着不同的效果。随着网络层数的增加，层数过深的网络会导致精度下降，模型效果甚至会变差，这与梯度在传播过程中梯度逐渐消失导致无法对网络权重进行迭代变换有关。因此，为了避免网络结构过深导致模型效果变差，适当地调整网络结构堆叠能够让模型充分发挥效果。本文的 CNN 网络结构是在 Piczak [21]和 Zhang [26]提出的网络结构的基础上调整提取特征的网络结构，调整它们的堆叠方式，改变卷积层和池化层的数量，以更好的对音频数据进行高层次特征提取。

模型结构构造为如图 3 所示，包括：一层输入层，堆叠五层卷积层和池化层，两层全连接层以及一个 softmax 分类层。此处的 softmax 分类层是用于训练网络时传递误差，当网络训练完成后，去掉该层而提取前一层的输出数据以进行下一步的训练。由于本文 CNN 网络只改变网络层数量，而不改变相关参数，因此参数设置与 Piczak 和 Zhang 所提出的网络参数设置一致，卷积核使用 3 × 3 矩阵，池化层步长使用 2 × 2 的最大池化方法，激活函数使用 relu 函数，通过该模型训练音频数据，提取音频文件的高层次特征数据。模型的主要流程如下：

算法：CNN 提取高层次特征

输入：音频文件经过预处理的 MFCC 矩阵

输出：音频文件高层次特征数据

1) 构建卷积神经网络模型并且初始化权值；

2) 输入数据经过卷积层、池化层、全连接层的前向传播得到输出值；

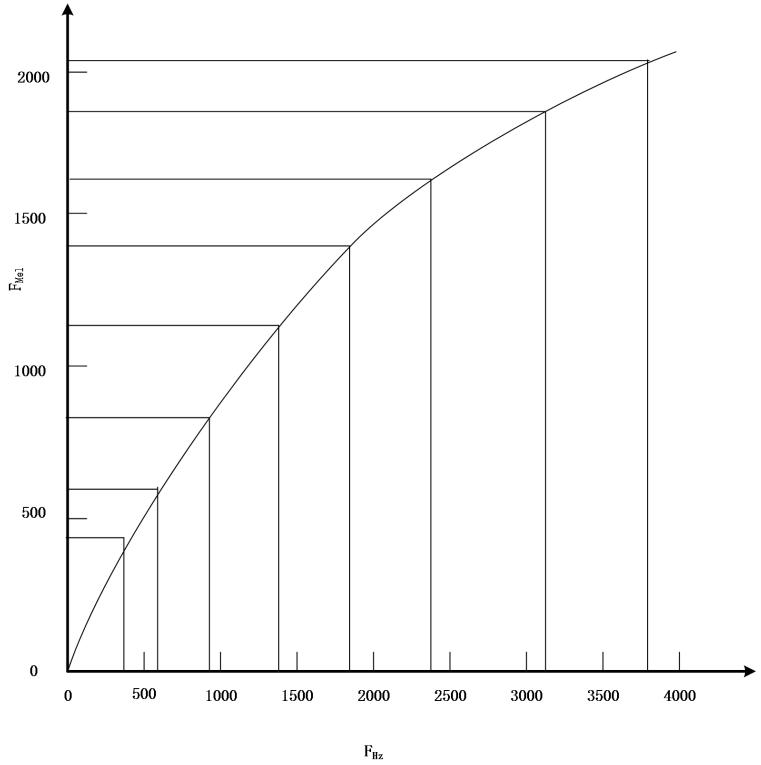

Figure 2. Frequency mapping to Mel frequency
图 2. 频率映射为 Mel 频率

3) 计算网络的输出值与目标值之间的误差；

4) 当误差大于期望值时，将误差传回网络中，依次求得网络全连接层、池化层、卷积层的误差，当误差小于等于给定期望值时，结束训练，进入第 6 步；

5) 根据求得误差调整网络各层权值，再次进入到第 2 步；

6) 训练结束，输出高层次特征数据。

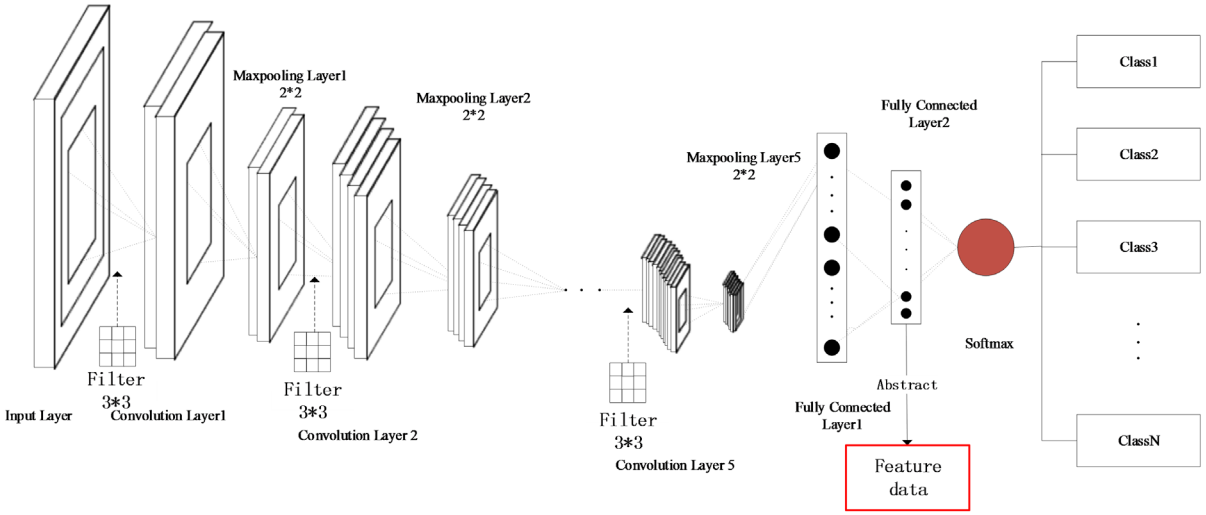

Figure 3. CNN extracts high-level feature model structure
图 3. 卷积神经网络提取高层次特征模型结构图

2.3. Light GBM

卷积神经网络的 softmax 层虽然也能进行分类预测，但是让模型既进行提取特征也进行分类会让模型泛化能力不强，因此考虑让 CNN 模型进行单独提取特征功能，而使用另一个分类器模型进行分类预测。

轻量级梯度提升器(Light Gradient Boosting Machine, Light GBM) [27]是集成学习中的一种 Boosting 框架，使用决策树作为学习算法的基分类器。Boosting 方法训练基分类器时采用串行训练的方式，各分类器之间有依赖。它的基本思想是将各个基分类器层层叠加，将弱分类器提升为强分类器，每一次训练时，对前面训练错误的样本赋予更高的权重，逐渐堆叠分类器为一个复杂强大的集成分类器。预测时，根据各层分类器结果加权取得最终预测结果。基于 Boosting 思想而出现的一系列算法模型，从早期的简单弱分类器组合成强分类器的自适应提升器(Adaptive Boosting, AdaBoost)到依据模型损失函数的梯度信息来迭代训练的梯度提升决策树(Gradient Boosting Decent Tree, GBDT)，再到在各大算法竞技平台普遍使用的 XgBoost (eXtreme Gradient Boosting)，以及在 XgBoost 基础上改进的 Light GBM 算法，这些算法模型在处理常见的分类预测问题上均能表现出不错的效果。算法的主要流程如下：

算法：Light GBM 分类预测

输入：经过 CNN 提取的高层次音频特征数据

输出：音频类别概率矩阵

1) 初始化 m 棵基分类决策树，训练样例的权重为$1/m$；

2) 训练弱分类器 $f(x)$；

3) 根据训练误差确定当前弱分类器 $f(x)$ 的权重 ∂；

4) 当未达到最大迭代次数，返回第 2 步继续训练，当达到最大迭代次数，进入第 5 步；

5) 得到最终分类器如公式(2)：

$$f_m(x)=\partial_0 f_0(x)+\partial_1 f_1(x)+\partial_2 f_2(x) \\ +\cdots+\partial_i f_i(x)+\cdots+\partial_m f_m(x) \tag{2}$$

6) 合并基分类器为强分类器对结果进行预测。

其中，m 为算法迭代次数，i 为第 i 代迭代，且 $0\le i\le m$，为第 i 代训练出的分类器。

2.4. 模型描述

针对单一的卷积神经网络模型在环境声音分类预测应用中功能单一导致的准确率不足问题，本文提出了融合两种模型的综合算法模型，具体改进内容如图 4 所示，在调整 CNN 网络结构的基础上，将原模型中的 softmax 方法改进为使用 Light GBM 方法进行分类。模型总共由三部分组成，分别为数据预处理模块、CNN 提取特征模块以及 Light GBM 分类预测模块。模型总体结构图如图 5 所示，先将原始音频数据通过数据预处理模块获取音频的物理特征 MFCC 矩阵然后通过 CNN 模块提取音频高层次特征数据，最后通过 Light GBM 模块训练进行分类预测得出预测结果。

算法模型的主要流程如下：

算法：基于 CNN 和 Light GBM 的环境声音分类

输入：音频文件信息

输出：模型分类预测概率矩阵

1) 通过数据预处理提取音频数据文件 MFCC 特征 T；

2) 构建卷积神经网络模型并且初始化权值；

3) 输入特征数据 T 经过卷积层、池化层、全连接层的前向传播得到输出值；

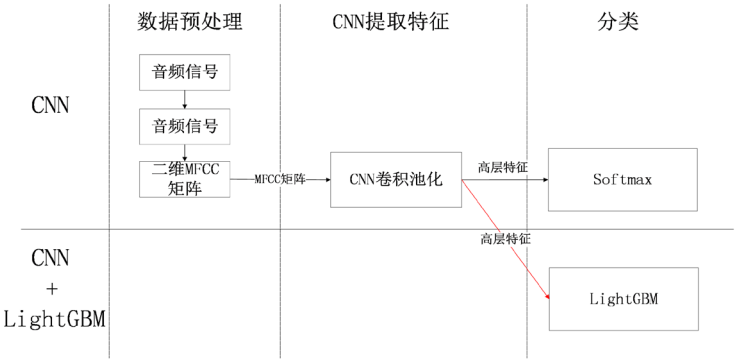

Figure 4. Classification method improvement point
图 4. 分类方法改进点

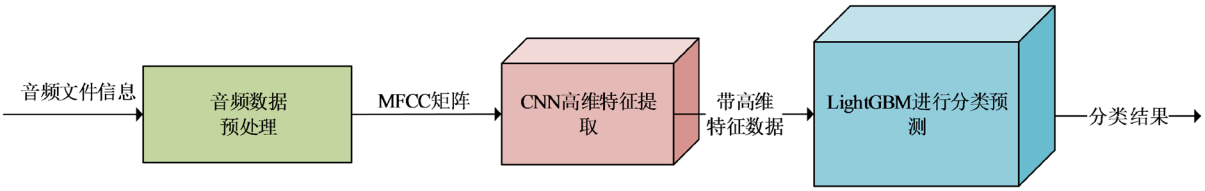

Figure 5. Model overall structure
图 5. 模型总体结构图

4) 计算网络的输出值与目标值之间的误差 $\text{error}=\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\widehat{y}_i\right)^2$；

5) 当误差大于期望值时，将误差传回网络中，依次求得网络全连接层、池化层、卷积层的误差，当误差大于给定期望值时，根据求得误差调整网络各层权值，进入到第 3 步，当误差小于等于给定期望值时，进入第 6 步；

6) 提取卷积神经网络 softmax 分类层前一层输出结果作为输入数据 V；

7) 初始化 Light GBM 基分类决策树

8) 训练弱分类器 $f(x)$；

9) 根据训练误差确定当前弱分类器 $f(x)$ 的权重 ∂；

10) 当未达到最大迭代次数，返回第 8 步继续训练，当达到最大迭代次数，进入第 10 步；

11) 合并基分类器为强分类器 $f_m(x)$

12) 使用综合模型对测试数据进行分类预测

其中提取音频 MFCC 物理特征过程如图 1 所示，将音频文件信息使用统一标准提取为规定长度的二维 MFCC 矩阵特征。i 为第 i 个样本，N 为样本数，y_i 为第 i 个样本的真实值，y_i 为第 i 个样本的预测值，通过构建如图 3 所示卷积神经网络结构，初始化网络卷积核、全连接层、激活函数各参数等，使用二维 MFCC 矩阵训练参数并提取最终训练结果，不做最后 softmax 层的分类而是提取前一层具有音频高层次特征的一维数据。构建 LightGBM 模型并使用高层次特征数据进行训练，最终使用整体模型进行分类预测。

3. 实验及结果

3.1. 数据集描述

以监督学习方式训练深度神经网络结构主要问题之一是有效学习所需的计算量和含标签数据量。虽

然前者在某方面通过硬件改进和 GPU 计算在通用基础上得到解决，但后者非常依赖于相关领域知识，往往很难获得有效的源数据。然而，无论是在数量上还是记录大小上，公开可用的环境声音记录数据集都十分有限。考虑到手动注释的高成本，虽然随着新的录制设备的引入逐渐改善了这种情况，但它仍然是该领域新数据密集型方法发展的主要障碍之一。因为监督的深度学习模型的性能受可用于学习的数据集大小的影响强烈，所以选择一个合适的数据集对模型训练的性能影响很大。

本实验数据集使用公开数据集 UrbanSound8K [28]，这个数据集包含了 8732 个短音频样本，每个音频文件不超过 4 秒(有少量文件略大于 4 秒)，这些音频样本被预编为 10 类声音(空调发动机(Air conditioner, AI)、汽车鸣笛声(Car horn, CA)、小孩子玩闹声(Children playing, CH)、狗叫声(Dog barking, DO)、钻孔声(Drilling, DR)、发动机空转声(Engine idling, EN)、枪声(Gun shot, GU)、手提钻(Jackhammer, JA)、警报器声(Siren, SI)和街边音乐(Street music, ST))。我们对这些音频文件使用相同的处理方式将数据统一基准并消除数据分割对实验精度的影响。

3.2. 实验设置

3.2.1. 数据预处理

所使用数据集的所有音频文件采样率均为 22050 Hz 和 16 位字深 wav 格式的单声道信道，使用 1024 的 FFT 窗口和 50%的重叠以及它们的增量(沿时间维度谱图的一阶导数)将文件转换为频谱图。对所有的数据文件将其提取为 4 s 长度的频谱，超过 4 s 的文件截取为 4 s，不足 4 s 的文件通过填充 0 补足为 4 s。在时频变换后提取分段，以提取音频文件的梅尔频率倒谱系数，为了更好地利用数据以及优化模型训练，使用 z-score 标准化对所有训练文件进行标准化特征，并且同时作用于验证集和测试集上，以达到将音频文件转化为矩阵信息的操作，进而将矩阵输入模型进行训练。图 6 为某个音频文件经过分帧、加窗、FFT 变换得到的频谱图，将该频谱图转换为灰度图，以提取该音频文件 MFCC 特征的输出，输出结果为 40 × 173 的二维 MFCC 矩阵，样本数据集 D(8732, 40, 173)，将总数据集划分为训练集和测试集，其中 70%作为训练集，30%作为测试集，最终得到训练集 T(6112, 40, 173)，测试集 V(2620, 40, 173)。

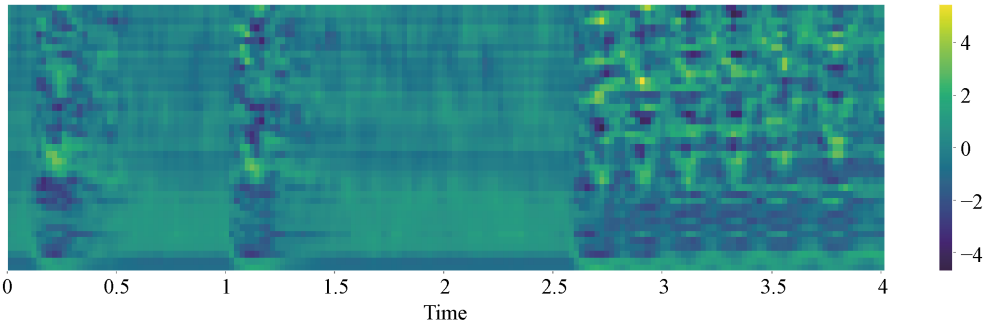

Figure 6. A Selection result of attracting audio feature
图 6. 一段音频文件提取特征信息的结果

3.2.2. 模型设置

为了提取频谱图的高层次特征，本文调整卷积神经网络层结构并利用数据集 D(8732, 40, 173)训练网络模型参数。由于频谱图为单通道数据，因此网络输入层大小设置为(40, 173, 1)，卷积层使用 3 × 3 的卷积核，初始使用 32 个过滤器，激活函数使用 relu 函数，卷积处理后填充边界，池化层使用 2 × 2 的核大小，步长设置为 2 × 2，堆叠 5 层卷积层和池化层后映射到全连接层再到 softmax 层，对应分类预测的结果，由于本实验数据共有 10 类，在 softmax 设置为 10 类。本实验所设置的卷积网络层结构具体见表 1，

将该结构卷积神经网络模型命名为 ProposedCNN。

Table 1. ProposedCNN related settings
表 1. 本文卷积神经网络层结构 ProposedCNN 相关设置

Layer	Ksize	Stride	Nums of filters	Out shape
Input	-	-	-	(40, 173, 1)
Conv1	(3, 3)	(1, 1)	32	(40, 173, 32)
Pool1	(2, 2)	(2, 2)	-	(20.86, 32)
Conv2	(3, 3)	(1, 1)	64	(20.86, 64)
Pool2	(2, 2)	(2, 2)	-	(10, 43, 64)
Conv3	(3, 3)	(1, 1)	128	(10, 43, 128)
Pool3	(2, 2)	(2, 2)	-	(5, 21, 128)
Conv4	(3, 3)	(1, 1)	256	(5, 21, 256)
Pool4	(2, 2)	(2, 2)	-	(2, 10, 256)
Conv5	(3, 3)	(1, 1)	512	(2, 10, 512)
Pool5	(2, 2)	(2, 2)	-	(1, 5, 512)
FC	-	-	2560	(2560)
Dense (Softmax)	-	-	Nums of classes	(10)

当卷积神经网络训练完成后，提取最后 softmax 层前一层的输出结果作为 LightGBM 模块的输入。LightGBM 模块中，输入训练数据由原来的 T(6112, 40, 173)重构为 T’(6112, 2560)，测试数据由原来的 V(2620, 40, 173)重构为 V’(2620, 2560)，并在训练过程中使用 GBDT 的提升类型，为了防止模型训练过程过拟合，设置较小的叶子节点数 num_leaves = 15、较小的学习率 learning_rate = 0.01 及较小的树深度 max_depth = 10。

实验将通过对比本文提出的 CNN 模型结构 ProposedCNN 和 Piczak 提出的卷积网络层结构框架 PiczakCNN、Zhang 提出的卷积网络层结构框架 ZhangCNN 训练过程的模型效果，其相关网络设置如表 2 和表 3 所示，同时比较本文提出 CNN 模型结构下融合不同分类学习方法的综合模型效果。

Table 2. PiczakCNN related settings
表 2. PiczakCNN 相关设置

Layer	Ksize	Stride	Nums of filters	Out shape
Input	-	-	-	(40, 173, 1)
Conv1	(3, 3)	(1, 1)	32	(40, 173, 32)
Pool1	(2, 2)	(2, 2)	-	(20.86, 32)
Conv2	(3, 3)	(1, 1)	64	(20.86, 64)
Pool2	(2, 2)	(2, 2)	-	(10, 43, 64)
Dropout	-	-	-	(10, 43, 64)
FC	-	-	2560	(2560)
Dense (Softmax)	-	-	Nums of classes	(10)

Table 3. ZhangCNN related settings
表 3. ZhangCNN 相关设置

Layer	Ksize	Stride	Nums of filters	Out shape
Input	-	-	-	(40, 173, 1)
Conv1-1	(3, 3)	(1, 1)	32	(40, 173, 32)
Conv1-2	(3, 3)	(1, 1)	32	(40, 173, 32)
Pool1	(2, 2)	(2, 2)	-	(20.86, 32)
Conv2-1	(3, 3)	(1, 1)	64	(20.86, 64)
Conv2-2	(3, 3)	(1, 1)	64	(20.86, 64)
Pool2	(2, 2)	(2, 2)	-	(10, 43, 64)
Conv3-1	(3, 3)	(1, 1)	128	(10, 43, 128)
Conv3-2	(3, 3)	(1, 1)	128	(10, 43, 128)
Pool3	(2, 2)	(2, 2)	-	(5, 21, 128)
Conv4-1	(3, 3)	(1, 1)	256	(5, 21, 256)
Conv4-2	(3, 3)	(1, 1)	256	(5, 21, 256)
Pool4	(2, 2)	(2, 2)	-	(2, 10, 256)
FC	-	-	2560	(2560)
Dense (Softmax)	-	-	Nums of classes	(10)

3.3. 实验结果及分析

在实验过程中，通过对比 PiczakCNN、ZhangCNN 和本文 ProposedCNN 训练过程，其训练过程中的模型准确率和模型损失如图 7、图 8 所示，从图中可以看出，不管是在训练集中还是在测试集中，本文提出的混合框架的模型准确率及模型损失均优于前面两种框架，通过对比表 4 中实验结果，进一步表明本文提出的 CNN 结构框架是有效的。

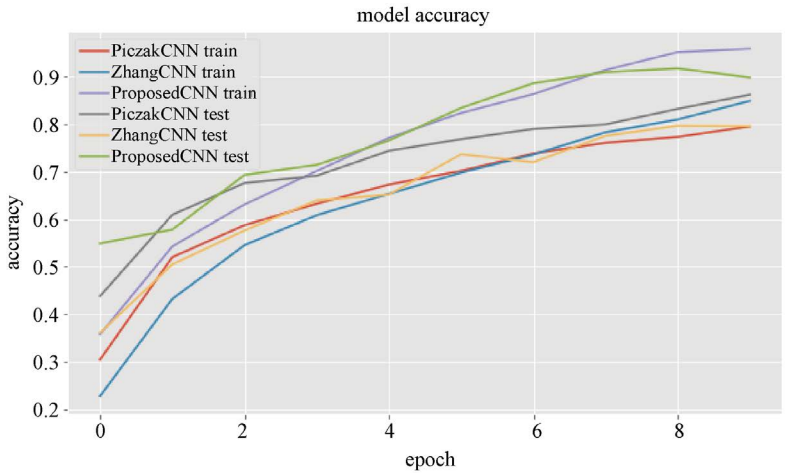

Figure 7. Accuracy of model training process under different convolutional layer structures
图 7. 不同卷积层结构下的模型训练过程准确率

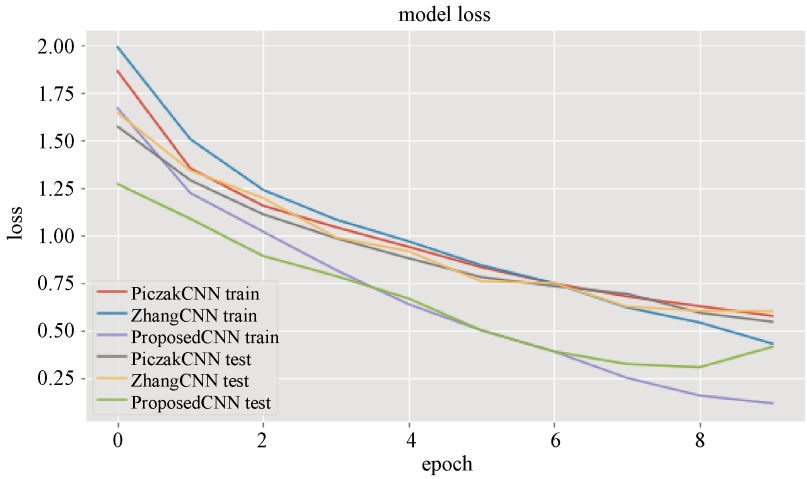

Figure 8. Loss of model training process under different convolutional layer
图 8. 不同卷积层结构下的模型训练过程损失

Table 4. Accuracy of different CNN structure
表 4. 不同 CNN 结构分类的准确率

Model	Accuracy
PiczakCNN	0.6618320608867033
ZhangCNN	0.7141221376775785
ProposedCNN	0.7519083969465649

通过对比本文所提出 CNN 结构结合不同分类方法实验比较，表 5 显示了各种模型的模型准确率和模型损失情况。从表中可以看出，实验的结果表明了融合 LightGBM 的综合模型，效果更优于融合其它常见的分类算法的混合模型，进一步证明了本文提出的混合结构框架的有效性，这说明融合的模型具有更好的预测分类效果，算法模型能够更好地应用于环境声音分类问题中。使用 ProposedCNN + LightGBM 混淆模型进行预测分类混合矩阵结果情况如图 9 所示，混淆矩阵是通过统计每个音频信息其对应的真实类别及预测类别的情况，矩阵中的数为统计真实类别样本被预测为指定类别的样本数，矩阵对角线上的数代表了正确分类预测到该类别的样本数，从混淆矩阵分布情况可以看出，大多数音频文件能够正确分类到对应类别中。

Table 5. Accuracy of ProposedCNN combined with different classification mechod
表 5. 本文所提 CNN 结构结合不同分类方法的准确率

Model	Accuracy
ProposedCNN (softmax)	0.7519083969465649
ProposedCNN + KNN	0.7660305346241434
ProposedCNN + RandomForest	0.7751908396946565
ProposedCNN + SVM	0.7851908396946565
ProposedCNN + LightGBM	0.7916030534351145

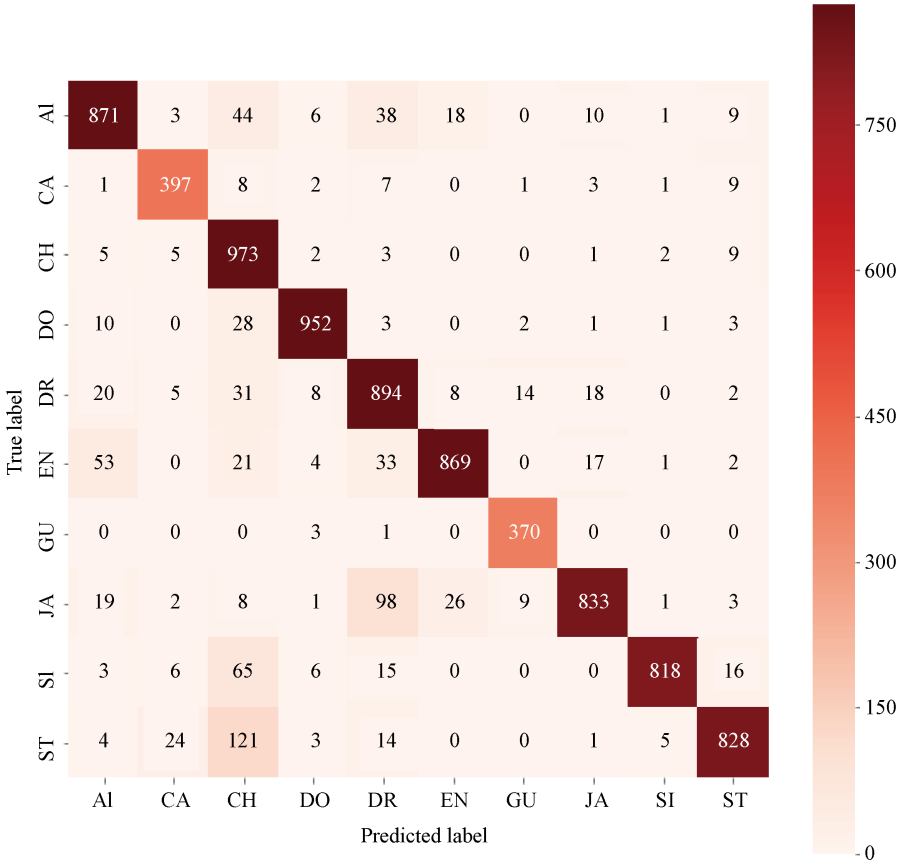

Figure 9. Model classification result confusion matrix
图 9. 模型分类结果混淆矩阵

4. 总结与展望

本文提出了一个新的将深度 CNN 与 LightGBM 融合的模型，并与过往提出过的网络结构模型进行对比，结果表明新的卷积神经网络结构模型具有更好的分类预测效果。综合模型通过提取音频文件数据，利用 CNN 提取音频的高维特征进行进一步的模型训练，通过对比实验表明使用综合模型可以进一步提高分类预测的准确性。

由于卷积网络结构具有不确定性，不能确定该结构网络模型是否是最佳的网络层结构，未来工作可以进一步探索具有更好效果的网络层结构模型，并且如果可以通过理论证明某个网络层结构模型在该工作中是具有最好效果的模型，将能给领域带来良好的贡献价值。

基金项目

本文得到国家自然科学基金项目(No.61572144)的资助。

本文得到广东省科技计划项目(No.2017B030307002，2015B010110001，2016B030306002)的资助。

参考文献

[1] Radhakrishnan, R., Divakaran, A. and Smaragdis, P. (2005) Audio Analysis for Surveillance Applications. 2005 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 16-19 October 2005, 1-4.

[2] Vacher, M., Serignat, J.F. and Chaillol, S. (2014) Sound Classification in a Smart Room Environment: An Approach Using GMM and HMM Methods. *The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue*,

Iasi, Romania, May 2007, 135-146.

[3] Barchiesi, D., Giannoulis, D., Stowell, D. and Plumbley, M.D. (2015) Acoustic Scene Classification: Classifying Environments from the Sounds They Produce. *IEEE Signal Processing Magazine*, **32**, 16-34. https://doi.org/10.1109/MSP.2014.2326181

[4] Lyon, R.F. (2010) Machine Hearing: An Emerging Field [Exploratory DSP]. *Signal Processing Magazine IEEE*, **27**, 131-139. https://doi.org/10.1109/MSP.2010.937498

[5] Kim, K. and Ko, H. (2011) Hierarchical Approach for Abnormal Acoustic Event Classification in an Elevator. 2011 8*th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Klagenfurt, Austria, 30 August-2 September 2011, 89-94. https://doi.org/10.1109/AVSS.2011.6027300

[6] Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T. and Okuno, H.G. (2011) Environmental Sound Recognition for Robot Audition Using Matching-Pursuit. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K. and Ali, M., Eds., *Modern Approaches in Applied Intelligence*. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 1-10. https://doi.org/10.1007/978-3-642-21827-9_1

[7] Eronen, A.J., Peltonen, V.T., Tuomi, J.T., *et al*. (2006) Audio-Based Context Recognition. *IEEE Transactions on Audio Speech & Language Processing*, **14**, 321-329. https://doi.org/10.1109/TSA.2005.854103

[8] Lee, K. and Ellis, D.P.W. (2010) Audio-Based Semantic Concept Classification for Consumer Video. *IEEE Transactions on Audio, Speech and Language Processing*, **18**, 1406-1416. https://doi.org/10.1109/TASL.2009.2034776

[9] Mcloughlin, I.V. (2008) Line Spectral Pairs. *Signal Processing*, **88**, 448-467. https://doi.org/10.1016/j.sigpro.2007.09.003

[10] Chu, S., Narayanan, S. and Kuo, C.C.J. (2009) Environmental Sound Recognition with Time-Frequency Audio Features. *IEEE Transactions on Audio, Speech and Language Processing*, **17**, 1142-1158. https://doi.org/10.1109/TASL.2009.2017438

[11] Valero, X. and Alias, F. (2012) Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. *IEEE Transactions on Multimedia*, **14**, 1684-1689. https://doi.org/10.1109/TMM.2012.2199972

[12] Geiger, J.T. and Helwani, K. (2015) Improving Event Detection for Audio Surveillance Using Gabor Filterbank Features. 2015 23*rd European Signal Processing Conference*, Nice, France, 31 August-4 September 2015, 714-718. https://doi.org/10.1109/EUSIPCO.2015.7362476

[13] 王熙, 李应. 多频带谱减法用于生态环境声音分类[J]. 计算机工程与应用, 2014, 50(3): 190-193.

[14] Temko, A., Monte, E. and Nadeu, C. (2005) Comparison of Sequence Discriminant Support Vector Machines for Acoustic Event Classification. 2006 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 14-19 May 2006, 5.

[15] Gupta, S., Dileep, A.D. and Thenkanidiyoor, V. (2016) Segment-Level Pyramid Match Kernels for the Classification of Varying Length Patterns of Speech Using SVMs. 2016 24*th European Signal Processing Conference*, Budapest, Hungary, 29 August-2 September 2016, 2030-2034. https://doi.org/10.1109/EUSIPCO.2016.7760605

[16] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M. and Plumbley, M.D. (2015) Detection and Classification of Acoustic Scenes and Events. *IEEE Transactions on Multimedia*, **17**, 1733-1746. https://doi.org/10.1109/TMM.2015.2428998

[17] Piczak, K.J. (2015) ESC: Dataset for Environmental Sound Classification. *IEEE Transactions on Wireless Communications*, **9**, 1015-1018.

[18] Hinton, G., Deng, L., Yu, D., *et al*. (2012) Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, **29**, 82-97. https://doi.org/10.1109/MSP.2012.2205597

[19] Dave, K. and Varma, V. (2014) Music Information Retrieval: Recent Developments and Applications. Now Publishers Inc., Hanover, MA. https://doi.org/10.1561/1500000045

[20] Mcloughlin, I., Zhang, H., Xie, Z., Song, Y. and Xiao, W. (2015) Robust Sound Event Classification Using Deep Neural Networks. *IEEE/ACM Transactions on Audio Speech & Language Processing*, **23**, 540-552. https://doi.org/10.1109/TASLP.2015.2389618

[21] Piczak, K.J. (2015) Environmental Sound Classification with Convolutional Neural Networks. 2015 *IEEE* 25*th International Workshop on Machine Learning for Signal Processing*, Boston, MA, 17-20 September 2015, 1-6. https://doi.org/10.1109/MLSP.2015.7324337

[22] Medhat, F., Chesmore, D. and Robinson, J. (2018) Masked Conditional Neural Networks for Environmental Sound Classification. 2017 *IEEE International Conference on Data Science and Advanced Analytics*, Tokyo, 19-21 October 2017, 389-394. https://doi.org/10.1109/DSAA.2017.43

[23] Takahashi, N., Gygli, M., Pfister, B. and Van Gool, L. (2016) Deep Convolutional Neural Networks and Data Aug-

mentation for Acoustic Event Detection. *Proceedings of Interspeech* 2016, 2982-2986. https://doi.org/10.21437/Interspeech.2016-805

[24] Zhang, H., Mcloughlin, I. and Song, Y. (2015) Robust Sound Event Recognition Using Convolutional Neural Networks. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, 19-24 April 2015, 559-563. https://doi.org/10.1109/ICASSP.2015.7178031

[25] Zhang, X., Zou, Y. and Shi, W. (2017) Dilated Convolution Neural Network with LeakyReLU for Environmental Sound Classification. 2017 22nd International Conference on Digital Signal Processing, London, 23-25 August 2017, 1-5. https://doi.org/10.1109/ICDSP.2017.8096153

[26] Zhang, Z., Xu, S., Cao, S. and Zhang, S. (2018) Deep Convolutional Neural Network with Mixup for Environmental Sound Classification. In: Lai, J.H., *et al*., Eds., *Pattern Recognition and Computer Vision*. *Lecture Notes in Computer Science*, Springer, Cham, 356-367. https://doi.org/10.1007/978-3-030-03335-4_31

[27] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, CA, 4-9 December 2017, 3146-3154.

[28] Salamon, J., Jacoby, C. and Bello, J.P. (2014) A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the* 22*nd ACM international conference on Multimedia*, Orlando, FL, 3-7 November 2014, 1041-1044. https://doi.org/10.1145/2647868.2655045