文章编号:1001-506X(2022)05-1433-06 网址:www. sys-ele. com

基于 NMF 与 CNN 联合优化的声学场景分类

韦 娟[1,*],杨皇卫[1],宁方立[2]
(1. 西安电子科技大学通信工程学院,陕西 西安 710071;2. 西北工业大学机电学院,陕西 西安 710072)

摘 要:针对声学场景分类任务中复杂声学环境的特征表示问题,提出一种联合训练特征提取和分类模型的优化算法。将非负矩阵分解与卷积神经网络的训练相结合,利用网络的损失值实现对特征提取和网络参数的共同更新,以学习到更具判别性的有监督特征。在 TUT2017 数据集上提取对数声谱图作为基础特征,搭建深度卷积神经网络进行实验验证。仿真结果表明,所提算法的识别准确率相比优化前提升 3.9%,且优于其他两种常用声学特征,证明该算法能够有效提升整体分类效果。

关键词:特征学习;非负矩阵分解;卷积神经网络;联合优化
中图分类号:O 42 **文献标志码**:A **DOI**:10.12305/j.issn.1001-506X.2022.05.01

Acoustic scene classification based on joint optimization of NMF and CNN

WEI Juan[1,*], YANG Huangwei[1], NING Fangli[2]
(1. *School of Communication Engineering, Xidian University, Xi'an 710071, China*;
2. *School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China*)

Abstract: To solve the problem of feature representation of complex acoustic environment in acoustic scene classification task, an optimization algorithm of joint training feature extraction and classification model is proposed. In order to learn more discriminative and supervised features, non-negative matrix factorization is combined with convolution neural network training, and the loss value of network is used to realize feature extraction and network parameters updating. The logarithmic spectrogram is extracted from the TUT2017 dataset as the basic feature. And the deep convolutional neural network is built for experimental verification. The simulation results show that the recognition accuracy of the proposed algorithm is improved by 3.9% compared with that before optimization, and is superior to the other two commonly used acoustic features, which proves that the algorithm can effectively improve the overall classification effect.

Keywords: feature learning; non-negative matrix factorization; convolutional neural network; joint optimization

0 引 言

声学场景分类(acoustic scene classification, ASC)旨在从不同音频片段中识别出各自包含的场景信息并加以分类。相比利用图像或视频信息实现场景分类,ASC 技术具有全向性,且不会受遮挡和光线条件的影响,在智能穿戴设备、物联网音频监控、巡检机器人等领域有着广泛的应用前景[1-2]。

实际声学场景通常由多个声学事件组成,但只有少数声学事件能对场景分析起到关键作用,因此需要提取足够有效的声学特征。ASC 任务中常用的对数梅尔谱(log Mel-spectrogram,LM)[3-4]和常数 Q 变换(constant Q transform, CQT)[5]可以对频带相对固定的音频信号进行有效的时频分析,但对于结构性较差的声学场景信号表现不佳[6]。于是,基于自动特征学习的非负矩阵分解(non-negative matrix factorization,NMF)[7-8]被应用于 ASC 任务。作为一种基于部分表达整体的方法,NMF 能够有效解决由各类声学事件组成的场景分类问题。姚琨等人[9]将 NMF 与 LM 进行特征融合以提高识别率,但未考虑样本标签对特征提取的辅助作用。Lee 等人[10]提出一种利用标签信息对各类声学场景独立学习基矩阵的方法,但不同场景可能存在相似的

收稿日期:2021-05-28;**修回日期**:2021-09-26;**网络优先出版日期**:2021-10-27。
网络优先出版地址:https://kns.cnki.net/kcms/detail/11.2422.TN.20211027.1935.012.html
基金项目:国家自然科学基金(52075441);陕西省重点研发计划(2018GY-181,2020ZDLGY06-09)资助课题
*通讯作者.
引用格式:韦娟,杨皇卫,宁方立. 基于 NMF 与 CNN 联合优化的声学场景分类[J]. 系统工程与电子技术,2022,44(5):1433-1438.
Reference format: WEI J, YANG H W, NING F L. Acoustic scene classification based on joint optimization of NMF and CNN[J]. Systems Engineering and Electronics, 2022, 44(5): 1433-1438.

声学事件，易造成基向量的冗余和混淆。Bisot 等人[11]提出基于逻辑回归的任务驱动型 NMF（task-driven NMF，TNMF）算法，通过分类器修正特征学习的方式有效提升场景分类效果，但因逻辑回归分类器性能有限而难以得到更有判别性的特征。

如何利用声学特征训练出有效的分类模型是 ASC 任务的另一个难点。随着深度学习的快速发展，卷积神经网络（convolutional neural network，CNN）[12]因为可以识别缩放、移位等空间失真不变性[13]，在 ASC 任务中得到广泛应用。Boddapati 等人[14]通过叠加声谱图、梅尔倒谱系数以及相干复原图组成三通道特征，结合图像识别中两种常用的 CNN 模型进行环境声音分类。Doan 等人[15]提出一种应用于耳蜗谱图的深度 CNN 模型，通过加深卷积层个数学习更丰富的场景信息。曹毅等人[16]将马尔可夫模型的思想应用于 CNN，提出一种更适合音频分类的 N 阶密集 CNN 模型。虽然上述模型尝试从不同角度获取特征图中的分类信息并取得一定的效果，但均基于一次性提取的无监督特征图，没有考虑在后续模型训练过程中对特征图本身所包含的信息进行优化。

针对以上问题，提出一种 NMF 与 CNN 联合优化的有监督特征学习算法。该算法利用基于 NMF 的特征表示训练 CNN 模型，根据标签信息和实际训练效果不断反向优化 NMF 的过程，自适应地调整特征提取方向以获得更利于分类的判别性特征。

1 特征提取

NMF 在对原始时频图降维的同时能够提取出声学场景的更好表示[17]。一方面，对非负声谱图矩阵 $\boldsymbol{V}$ 进行 NMF，可理解为联合学习非负的基矩阵 $\boldsymbol{W}$ 与权值矩阵 $\boldsymbol{H}$，使得 $\boldsymbol{V}\approx\boldsymbol{WH}$[18-19]。其中，$\boldsymbol{W}$ 的列向量代表特定的声学事件，$\boldsymbol{H}$ 的列向量对应当前时刻各声学事件所占的比重。由于声学场景是由不同声学事件组成的复杂多源环境，因此判断特定事件是否发生将有助于分辨不同的场景。另一方面，NMF 算法可以与标签信息结合，不断修正特征提取过程，促使基矩阵 $\boldsymbol{W}$ 对环境中声学事件的刻画更加准确，从而增强 NMF 特征的表达能力。

对音频样本进行短时傅里叶变换得到声谱图 $\boldsymbol{v}\in\mathbf{R}_{+}^{F\times T}$，其中 F 表示频带数，T 表示时间帧数。将所有训练集样本的声谱图扩展得到矩阵 $\boldsymbol{V}\in\mathbf{R}_{+}^{F\times NT}$，其中 N 表示训练集样本的总数。NMF 算法的目的是在给定矩阵 $\boldsymbol{V}$ 下，利用乘性更新法则找到基矩阵 $\boldsymbol{W}\in\mathbf{R}_{+}^{F\times K}$ 和权值矩阵 $\boldsymbol{H}\in\mathbf{R}_{+}^{K\times NT}$，$K\leqslant FNT/(F+NT)$。可表示为如下优化问题[6]：

$$\min_{\boldsymbol{W},\boldsymbol{H}} f(\boldsymbol{W},\boldsymbol{H})=\min_{\boldsymbol{W},\boldsymbol{H}}\frac{1}{2}\left(\|\boldsymbol{V}|\boldsymbol{WH}\|_{\mathrm{F}}^{2}+\frac{\lambda}{2}\|\boldsymbol{H}\|_{2}^{2}\right) \tag{1}$$

式中：$\|\cdot\|_{\mathrm{F}}$ 表示矩阵 Frobenius 范数；λ 表示 L_2 正则化系数，目的是防止基矩阵出现过拟合。

通过 NMF 算法得到基矩阵 $\boldsymbol{W}$，再对每个样本的声谱图 $\boldsymbol{v}$ 在 $\boldsymbol{W}$ 上利用带有正约束的最小角回归算法[20]进行投影，得到的权值矩阵 $\boldsymbol{h}$ 即为该样本的 NMF 特征。

进一步，令 $\partial f(\boldsymbol{W},\boldsymbol{h})/\partial\boldsymbol{h}=\boldsymbol{0}$，有：

$$\boldsymbol{h}=(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\lambda\boldsymbol{I})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{v} \tag{2}$$

对式(2)求微分，有：

$$\mathrm{d}\boldsymbol{h}=-(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\lambda\boldsymbol{I})^{-1}\boldsymbol{W}^{\mathrm{T}}\mathrm{d}\boldsymbol{W}\boldsymbol{h}+(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\lambda\boldsymbol{I})^{-1}(\mathrm{d}\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{v}-\boldsymbol{W}\boldsymbol{h}) \tag{3}$$

式(3)表达了权值矩阵 $\boldsymbol{h}$ 与基矩阵 $\boldsymbol{W}$ 的微分关系，利用该式以及样本的标签信息即可根据联合优化算法实现 NMF 特征的修正。

2 联合优化算法

NMF 作为一种自动特征学习方法，能够根据不同任务和数据集自动学习到有效特征。在加入标签信息后，NMF 可进一步调整特征提取方向，提高对特定任务的适应性。于是，在文献[11]的基础上提出一种联合优化算法，通过引入 CNN 模型实现 NMF 与神经网络的联合训练，提取同时包含生成信息和判别信息[21]的有监督 NMF（supervised NMF，SNMF）特征。

令神经网络的损失函数为 $\boldsymbol{l}_s$，有：

$$\mathrm{d}\boldsymbol{l}_s=\mathrm{tr}[\boldsymbol{\nabla}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{l}_s\mathrm{d}\boldsymbol{h}]=\mathrm{tr}[\boldsymbol{\nabla}_{\boldsymbol{W}}^{\mathrm{T}}\boldsymbol{l}_s\mathrm{d}\boldsymbol{W}] \tag{4}$$

式中：tr[·]代表矩阵的迹。类似于利用损失函数 $\boldsymbol{l}_s$ 对网络中各参数进行梯度反向修正的过程，$\boldsymbol{l}_s$ 对权值矩阵 $\boldsymbol{h}$ 的偏导数 $\nabla_{\boldsymbol{h}}\boldsymbol{l}_s$ 可由深度学习框架 Keras 中的函数 keras.backend.gradients()得到。结合式(3)中权值矩阵 $\boldsymbol{h}$ 与基矩阵 $\boldsymbol{W}$ 的微分关系，可得反向传播模型为

$$\boldsymbol{\nabla}_{\boldsymbol{W}}\boldsymbol{l}_s=-\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\nabla}_{\boldsymbol{h}}\boldsymbol{l}_s\boldsymbol{h}^{\mathrm{T}}+(\boldsymbol{v}-\boldsymbol{W}\boldsymbol{h})\boldsymbol{\nabla}_{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{l}_s(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\lambda\boldsymbol{I})^{-1} \tag{5}$$

利用梯度下降法实现基矩阵的修正：

$$\boldsymbol{W}=\Pi_{\boldsymbol{W}}[\boldsymbol{W}-\rho\boldsymbol{\nabla}_{\boldsymbol{W}}\boldsymbol{l}_s] \tag{6}$$

式中：$\Pi_{\boldsymbol{W}}$ 表示对基矩阵 $\boldsymbol{W}$ 进行 L_2 范数标准化；ρ 代表基矩阵的学习率。

在修正后的基矩阵上进行投影，得到新的权值矩阵 $\boldsymbol{h}$ 即为 SNMF 特征。

算法具体步骤如下。

步骤 1 将训练集样本的声谱图扩展后进行 NMF，得到基矩阵 $\boldsymbol{W}$。

步骤 2 将训练集样本的声谱图在基矩阵 $\boldsymbol{W}$ 上进行投影，获得的权值矩阵输入已搭建的 CNN 模型中进行训练。

步骤 3 从训练集中随机不重复选取一组样本的声谱图，在基矩阵 $\boldsymbol{W}$ 上投影得到权值矩阵 $\boldsymbol{h}$，输入已训练 CNN 模型中获取对应的一组损失值。

步骤 4 利用式(6)实现网络损失值对基矩阵 $\boldsymbol{W}$ 的修正。

步骤 5 在修正后的基矩阵 $\boldsymbol{W}$ 基础上重复步骤 3～步骤 4，完成整个训练集样本对基矩阵的修正。

步骤 6 在修正完毕的基矩阵 $\boldsymbol{W}$ 基础上重复步骤 2～步骤 5，直到满足预设条件后退出循环。

联合优化算法的整体流程如图 1 所示。

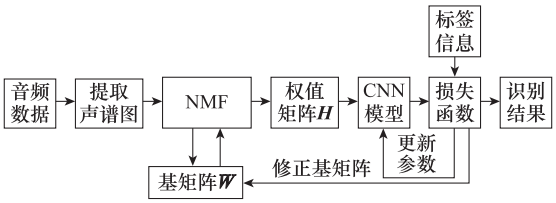

图 1　联合优化算法的流程框图

Fig. 1　Flow chart of joint optimization algorithm

3　网络模型

目前 ASC 任务主要采用 CNN 型深度神经网络对二维时频特征进行分类[22-23]。通过 NMF 得到的二维特征同样包含丰富的分类信息，可使用相似的网络结构[10]。在模型的搭建上，一方面，由于各时间片段里包含的声学事件有所不同，为使模型充分学习到这些声学事件的有效特征，应适当减少在时间轴上的池化。另一方面，网络中的卷积层数目对识别效果也有一定影响[24]。数目过少可能导致网络的拟合程度不高；数目过多则可能因梯度消失问题降低 SNMF 特征的修正效果。为得到适合 SNMF 特征的模型，并验证网络层数对分类效果的影响，在视觉几何组网络（visual geometry groupnet work，VGGNet）[25] 和文献[3]的基础上分别搭建卷积层数目为 8、10、12 的 CNN8、CNN10 和 CNN12 模型进行实验。

模型结构与参数如表 1 所示，其中@符号表示 Conv2D 卷积层。优化器使用随机梯度下降算法，批大小为 16，模型的训练与 SNMF 特征的修正交替进行。为避免因网络收敛过快而导致修正幅度较小，选择每训练 10 轮模型修正 1 次 SNMF 特征。每 10 轮间模型的学习率按热重启学习率策略[3,26]从 5×10^{-3} 以余弦下降方式衰减到 5×10^{-5}，使用交叉熵损失函数共训练 70 轮[11]。

表 1　CNN 模型结构

Table 1　CNN model structure

名称	CNN8	CNN10	CNN12
输入层	256×108×1	256×108×1	256×108×1
批归一化层，卷积层	BN，3×3@64	BN，3×3@64	BN，3×3@64
批归一化层，激活层，卷积层	BN，ReLu，3×3@64	BN，ReLu，3×3@64	BN，ReLu，3×3@64
池化层	4×2AvgPooling	4×2AvgPooling	4×2AvgPooling
批归一化层，激活层 卷积层	(BN，ReLu 3×3@128)×2	(BN，ReLu 3×3@128)×2	(BN，ReLu 3×3@128)×2
池化层	4×2AvgPooling	4×2AvgPooling	4×2AvgPooling
批归一化层，激活层 卷积层	(BN，ReLu 3×3@256)×2	(BN，ReLu 3×3@256)×2	(BN，ReLu 3×3@256)×2
池化层	—	2×1AvgPooling	2×1AvgPooling
批归一化层，激活层 卷积层	— —	(BN，ReLu 3×3@512)×2	(BN，ReLu 3×3@512)×2
池化层	—	—	2×1AvgPooling
批归一化层，激活层 卷积层	— —	— —	(BN，ReLu 3×3@1024)×2
批归一化层，激活层，卷积层	BN，ReLu，1×1@1024		
批归一化层，卷积层，全局池化层	BN，1×1@15，Global AvgPooling		
全连接层，输出层	Dense(15)，Softmax		

4　实验与分析

4.1　实验数据与配置

实验所采用的数据集为 TUT Acoustic Scenes 2017 开发数据集[27]。该数据集的录音时长总计 13 h，包括沙滩、公交、咖啡馆/饭馆、汽车、市中心、林荫道、杂货店、家、图书馆、地铁站、办公室、公园、居民区、火车、电车在内的 15 种声学环境，每类音频包含 312 个样本，总共 4 680 个样本。样本均为采样率 44.1 kHz，精度 24 位，时长 10 s 的双声道音频。将所有样本降采样到 22.05 kHz，平均左右声道数据以供后续使用。根据官方提供的四折交叉验证方式进行数据集的划分与实验，使用准确率作为最终的评价指标。实验硬件配置为 Intel(R) Core(TM) i5-10400F CPU、16 GB 内存、Nvidia GeForce RTX 2060 GPU，软件环境为 Ubuntu18.04 系统，Python3.6.11、Tensorflow1.15.0、Keras2.3.1。

4.2　参数设置

NMF 特征设置：帧长和帧移分别为 1 024 和 512 个采样点，通过短时傅里叶变换得到 512×431 的对数声谱图。按文献[11]的方法进行池化操作得到 512×108 的对数声谱图。扩展所有训练样本的声谱图后进行 NMF 得到 $512\times K$ 的基矩阵 $\boldsymbol{W}$，K 为基向量数及特征维数，该基矩阵同时用作 SNMF 特征的初始基矩阵。最后在 $\boldsymbol{W}$ 上重新投影得到 $K\times108$ 的 NMF 特征。

SNMF 特征设置：正则化系数 λ 设为 2×10^{-2}，学习率

ρ 取 5×10^{-4}。参数的选择来源于组合实验的结果。

其他特征设置：为了对比分析，提取由声谱图通过 256 组梅尔滤波器后获得的 LM 特征，尺寸为 256×431；每 8 度取 24 个频带得到的 CQT 特征，尺寸为 255×431。通过池化操作后得到 256×108 的 LM 特征与 255×108 的 CQT 特征。

4.3 结果分析

4.3.1 特征维数和模型层数对准确率的影响

为说明不同特征维数对分类准确率的影响，令分类器为已搭建的 CNN10 模型，并分别令 NMF 中基向量的数目为 64、128、256 和 512 以提取 4 种不同维数的 SNMF 特征。如表 2 所示，为 SNMF 特征在四折交叉验证下取不同特征维数时的准确率变化情况。

表 2 不同特征维数下的准确率

Table 2 Accuracy of different feature dimensions

SNMF	Fold1	Fold2	Fold3	Fold4	平均
$K=64$	0.781	0.795	0.771	0.824	0.793
$K=128$	0.805	0.837	0.793	0.854	0.822
$K=256$	0.827	0.839	0.814	0.863	0.836
$K=512$	0.818	0.831	0.807	0.855	0.828

由表 2 可知，K 值取 64、128 和 256 时，四折数据划分下的识别准确率均随着特征维数的增加而提高。说明随着基向量的增多，基矩阵对声学场景中各声学事件的学习更加充分，能够从声谱图中学习到更细分的基事件，使提取的 SNMF 特征中包含更多的区分信息。但当 K 值大于 256 时，识别准确率发生一定下降，说明 K 值并非越大越好。因为，此时多余的基向量学习到的是噪声和冗余信息，将对识别效果产生一定干扰。

表 3 为 $K=256$ 时 SNMF 特征在模型取不同层数时对识别准确率的影响。

表 3 不同模型层数下的准确率

Table 3 Accuracy of different model layers

模型	Fold1	Fold2	Fold3	Fold4	平均
CNN8	0.808	0.807	0.778	0.806	0.800
CNN10	0.827	0.839	0.814	0.863	0.836
CNN12	0.811	0.815	0.788	0.861	0.819

分析表 3 可知，模型的层数会对识别准确率产生较大影响。层数较低时，因为网络欠拟合而导致分类效果不佳；而层数较高时则容易因网络过深而产生梯度消失问题。由于联合优化算法的效果依赖于网络损失值的梯度反向传播，若出现梯度消失将会使 SNMF 特征的修正程度不高，从而降低联合优化算法的效果。

4.3.2 不同特征之间的对比

为验证联合优化算法的实际效果，将 SNMF 特征与 TUT2017 数据集的官方基线系统[27]、无监督 NMF 特征、以对数声谱图为基础提取的 TNMF 特征[11]、CQT 特征与 LM 特征进行对比。其中，NMF 特征和 SNMF 特征的特征维数 $K=256$。为保证所有特征能够拥有适合其自身特点的分类器，令 NMF 与 SNMF 特征的分类器为 CNN10 模型，TNMF 特征的分类器同文献[11]，而 LM 和 CQT 特征则选取在 2020 年声学场景和事件的检测与分类挑战赛（Detection and Classification of Acoustic Scenes and Events，DCASE）中表现优异的类 VGGNet 模型[28]。获得的分类结果如表 4 所示。

表 4 不同特征的识别准确率对比

Table 4 Comparison of recognition accuracy of different features

场景	基线系统	NMF	TNMF	SNMF	CQT	LM
沙滩	0.753	0.751	0.747	0.835	0.895	0.887
公交	0.718	0.893	0.813	0.928	0.930	0.922
饭馆	0.577	0.618	0.544	0.793	0.611	0.628
汽车	0.971	0.962	0.945	0.942	0.978	0.941
市中心	0.907	0.943	0.867	0.893	0.778	0.920
林荫道	0.795	0.769	0.892	0.925	0.881	0.855
杂货店	0.587	0.801	0.828	0.920	0.883	0.929
家	0.686	0.702	0.662	0.792	0.820	0.663
图书馆	0.571	0.725	0.691	0.658	0.783	0.685
地铁站	0.917	0.742	0.826	0.815	0.852	0.747
办公室	0.998	0.965	0.950	0.941	0.875	0.942
公园	0.702	0.695	0.712	0.705	0.545	0.723
居民区	0.641	0.874	0.774	0.738	0.691	0.764
火车	0.580	0.657	0.768	0.802	0.685	0.712
电车	0.817	0.852	0.847	0.851	0.864	0.876
总体	0.748	0.797	0.791	0.836	0.805	0.813
预测时间/s	—	2.6	1.1	2.7	3.1	3.3

分析表 4 可知，与 CNN 结合的无监督 NMF 特征和 SNMF 特征的识别准确率分别高出基线系统 4.9% 和 8.8%，说明 NMF 与 CNN 结合是一种有效的识别方法。同时，即使未使用联合优化算法的 NMF 特征也要优于使用逻辑回归分类器的 TNMF 特征，说明分类器的性能对识别结果有着较大影响。另外，通过联合优化算法获取的 SNMF 特征识别准确率达到 83.6%，分别高出 NMF 特征 3.9%、CQT 特征 3.1% 和 LM 特征 2.3%，说明联合优化算法有助于提取更优的特征。原因是与 CNN 分类器相结合的有监督特征学习方式能够利用标签信息和实际分类效果不断调整 NMF 中基矩阵内的参数，提高基向量的表征能力，从而获取更有判别性的特征。

由表 4 还可知，在不同类别场景下的分类效果方面，SNMF 特征在所有类别中准确率的最大值与最小值之间的差值最小，说明 SNMF 特征有更好的稳定性。另外，无论哪一种特征，在汽车、市中心、办公室、电车等类别的分类上均表现良好，而在某些类别的分类上性能却不高，如饭馆、图书馆、公园和居民区。这主要是因为噪声影响使其具有的特定声学事件变得模糊不清，或是该类声学场景中具有易与其他声学场景造成混淆的相似声学事件[29-30]。而在测试集样本的总预测时间方面，几种特征没有明显的区别，都能够满足一般场景下的实时性要求。

5 结　论

为解决 ASC 任务中特征提取与模型训练的联合优化问题，首先对声谱图进行 NMF，得到基矩阵和权值矩阵，然后搭建并训练 CNN 模型，根据训练结果反向更新基矩阵以获得修正的 SNMF 特征，实现一种 NMF 与 CNN 联合优化的有监督特征学习方法。得出结论如下：

（1）提高特征维数有利于学习更细分的基事件，但维数过高则会因噪声和冗余信息降低识别效果；

（2）由于联合优化算法依赖于梯度反向传播，过高的网络层数会引起梯度消失从而影响算法的优化效果；

（3）相较于直接使用 NMF 特征，联合优化后的 SNMF 特征能够使分类准确率得到明显提升；

（4）所提方法实现了特征提取与网络训练的联合优化，是一种有效的声学场景分类方法。

参考文献

[1] PASEDDULA C, GANGASHETTY S V. Late fusion framework for acoustic scene classification using LPCC, SCMC, and log-Mel band energies with deep neural networks[J]. Applied Acoustics, 2021, 172: 107568.

[2] 刘立芳，杨海霞，齐小刚. 基于线性判别分析的时频域特征提取算法[J]. 系统工程与电子技术，2019，41(10)：2184－2190.
LIU L F, YANG H X, QI X G. Time-frequency domain feature extraction algorithm based on linear discriminant analysis[J]. Systems Engineering and Electronics, 2019, 41(10): 2184－2190.

[3] MCDONNELL M D, GAO W. Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths[C]// Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[4] SONG H W, HAN J Q, DENG S W, et al. Acoustic scene classification by implicitly identifying distinct sound events[C]// Proc. of the Interspeech, 2019: 3860－3864.

[5] WANG M, WANG R, ZHANG X L, et al. Hybrid constant-Q transform based CNN ensemble for acoustic scene classification[C]// Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2019: 1511－1516.

[6] BISOT V, SERIZEL R, ESSID S, et al. Feature learning with matrix factorization applied to acoustic scene classification[J]. IEEE/ACM Trans. on Audio Speech & Language Processing, 2017, 25(6): 1216－1229.

[7] SPRECHMANN P, BRONSTEIN A M, SAPIRO G. Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement[C]// Proc. of the Hands-free Speech Communication and Microphone Arrays, 2014: 11－15.

[8] PODWINSKA Z, SOBIERAJ I, FAZENDA B M, et al. Acoustic event detection from weakly labeled data using auditory salience[C]// Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

[9] 姚琨，杨吉斌，张雄伟，等. 基于多分辨率时频特征融合的声学场景分类[J]. 声学技术，2020，39(4)：108－114.
YAO K, YANG J B, ZHANG X W, et al. Acoustic scene classification based on multi-resolution time-frequency feature fusion[J]. Acoustic Technology, 2020, 39(4): 108－114.

[10] LEE S, PANG H S. Feature extraction based on the non-negative matrix factorization of convolutional neural networks for monitoring domestic activity with acoustic signals[J]. IEEE Access, 2020, 8: 122384－122395.

[11] BISOT V, SERIZEL R, ESSID S, et al. Supervised non-negative matrix factorization for acoustic scene classification[C]// Proc. of the IEEE International Evaluation Campaign on Detection and Classification of Acousitc Scenes and Events, 2016.

[12] SALAMON J, BELLOJ P. Deep convolutional neural networks and data augmentation for environmental sound classification[J]. IEEE Signal Processing Letters, 2017, 24(3): 279－283.

[13] 杨浩聪，史创，李会勇. 保留立体声相位信息的声音场景分类系统[J]. 信号处理，2020，36(6)：871－878.
YANG H C, SHI C, LI H Y. Sound scene classification system preserving stereo phase information[J]. Signal Processing, 2020, 36(6): 871－878.

[14] BODDAPATI V, PETEF A, RASMUSSON J, et al. Classifying environmental sounds using image recognition networks[J]. Procedia Computer Science, 2017, 112: 2048－2056.

[15] DOAN T, NGUYEN H, NGO D T, et al. Acoustic scene classification using adeeper training method for convolution neural network[C]// Proc. of the International Symposium on Electrical and Electronics Engineering, 2019: 63－67.

[16] 曹毅，黄子龙，张威，等. N-DenseNet 的城市声音事件分类模型[J]. 西安电子科技大学学报，2019，46(6)：9－16，94.
CAO Y, HUANG Z L, ZHANG W, et al. Urban sound event classification model based on N-DenseNet[J]. Journal of Xidian University, 2019, 46(6): 9－16，94.

[17] 李伟，李硕. 理解数字声音——基于一般音频/环境声的计算机听觉综述[J]. 复旦学报(自然科学版)，2019，58(3)：269－313.
LI W, LI S. Understanding digital sound: a review of computer hearing based on general audio/ambient sound[J]. Journal of Fudan University (Natural Science Edition), 2019, 58(3): 269－313.

[18] KOMATSU T, SENDA Y, KONDO R. Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation[C]// Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2016: 2259－2263.

[19] GIANNOULIS P, POTAMIANOS G, MARAGOS P. Multichannel non-negative matrix factorization for overlapped acoustic event detection[C]// Proc. of the 26th European Signal Processing Conference, 2018: 857－861.

[20] MAIRAL J, BACH F, PONCE J. Task-driven dictionary learning[J]. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2012, 34(4): 791－804.

[21] RAKOTOMAMONJY A. Supervised representation learning for audio scene classification[J]. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2017, 25(6): 1253－1265.

[22] PHAM L, MCLOUGHLIN I, PHAN H, et al. A robust framework for acoustic scene classification[C]// Proc. of the Interspeech, 2019: 3634－3638.

[23] LI X Y, CHEBIYYAM V, KIRCHHOFF K. Multi-stream network with temporal attention for environmental sound classification[C]// Proc. of the Interspeech, 2019: 3604 - 3608.

[24] KONG Q, CAO Y, IQBAL T, et al. Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems[EB/OL]. [2021 - 05 - 28]. http://arxiv.org/abs/1904.03476v3.

[25] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale imagerecognition[EB/OL]. [2021 - 05 - 28]. http://arxiv.org/abs/1409.1556v6.

[26] MCDONNELL M D. Training wide residual networks for deployment using a single bit for each weight[EB/OL]. [2021 - 05 - 28]. http://arxiv.org/abs/1802.08530.

[27] MESAROS A, HEITTOLA T, DIMENT A, et al. DCASE 2017 Challenge setup: tasks,datasets and baseline system[C]// Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop, 2017: 85 - 92.

[28] WANG H L, ZOU Y X, CHONG D D. Acoustic scene classification with spectrogram processing strategies[C]// Pro. of the Detection and Classification of Acoustic Scenes and Events Workshop, 2020.

[29] WANG C, SANTOSO A, WANG J. Acoustic scene classification using self-determination convolutional neural network[C]// Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2017: 19 - 22.

[30] DANG A, VUT H, WANG J. Acoustic scene classification using convolutional neural networks and multi-scale multi-feature extraction[C]// Proc. of the IEEE International Conference on Consumer Electronics, 2018.

作者简介

韦　娟(1973—),女,教授,博士,主要研究方向为声源定位、音频识别。

杨皇卫(1997—),男,硕士研究生,主要研究方向为声场景分类。

宁方立(1974—),男,教授,博士,主要研究方向为声源定位。