第 41 卷 第 3 期
2022 年 5 月

应用声学
Journal of Applied Acoustics

Vol. 41, No. 3
May, 2022

◇ 研究报告 ◇

神经网络的声场景自动分类方法*

梁 腾[1] 姜文宗[1] 王 立[2] 刘宝弟[2] 王延江[2†]

(1 中国石油大学 (华东) 海洋与空间信息学院 青岛 266580)

(2 中国石油大学 (华东) 控制科学与工程学院 青岛 266580)

摘要： 声场景探察和自动分类能帮助人类制定应对特定环境的正确策略，具有重要的研究价值。随着卷积神经网络的发展，出现了许多基于卷积神经网络的声场景分类方法。其中时频卷积神经网络 (TS-CNN) 采用了时频注意力模块，是目前声场景分类效果最好的网络之一。为了在保持网络复杂度不变的前提下进一步提高网络的声场景分类性能，该文提出了一种基于协同学习的时频卷积神经网络模型 (TSCNN-CL)。具体地说，该文首先建立了基于同构结构的辅助分支参与网络的训练。其次，提出了一种基于 KL 散度的协同损失函数，实现了分支与主干的知识协同，最后，在测试过程中，为了不增加推理计算量，该文提出的模型只使用主干网络预测结果。在 ESC-10、ESC-50 和 UrbanSound8k 数据集的综合实验表明，该模型分类效果要优于 TS-CNN 模型以及当前大部分的主流方法。

关键词： 声场景分类；时频卷积神经网络；协同学习；声信号处理

中图法分类号： TP39 **文献标识码：** A **文章编号：** 1000-310X(2022)03-0373-08

DOI: 10.11684/j.issn.1000-310X.2022.03.006

Automatic classification of acoustic scene based on neural network

LIANG Teng[1] JIANG Wenzong[1] WANG Li[2] LIU Baodi[2] WANG Yanjiang[2]

(1 *College of Oceanography and Space Information, China University of Petroleum* (*East China*), *Qingdao* 266580, *China*)

(2 *College of Control Science and Engineering, China University of Petroleum* (*East China*), *Qingdao* 266580, *China*)

Abstract: Acoustic scene detection and automatic classification can help human beings to make correct strategies in specific environments, which indicates great research values. With the development of convolutional neural networks (CNN), a large number of CNN-based acoustic scene classification methods emerge. Especially, the temporal-spectral CNN (TS-CNN) which adapts the temporal-spectral attention module, is one of the best methods for the classification of acoustic scenes at present. In order to further improve the acoustic scene classification ability of the neural network without changing the complexity, in this paper, we proposed a new temporal-spectral CNN model which was based on the collaborative learning method (TSCNN-CL) More specifically, first, we established the auxiliary branches based on the isomorphism to participate in the network training. Second, we adopt a collaborative loss function based on KL divergence to realize the knowledge collaboration between the branches and the trunk. Finally, in the testing process, only the network trunk was used to predict the results, leading to the invariant amount of inference calculation. Comprehensive experiments on ESC-10, ESC-50, and UrbanSound8k datasets showed that the classification performance of TSCNN-CL model outperformed the TS-CNN model and even had compelling advantages in comparison with some other state-of-art models.

Keywords: Acoustic scene classification; Temporal-spectral convolutional neural network; Collaborative learning; Sound signal processing

---

2021-04-12 收稿; 2021-06-21 定稿

*国家自然科学基金项目 (62072468)

作者简介: 梁腾 (1996–), 男, 山东德州人, 硕士研究生, 研究方向: 智能信息处理。

† 通信作者 E-mail: yjwang@upc.edu.cn

0 引言

声场景是指人们的日常环境和周围发生的各种物理事件所产生的声音。如，繁忙的街道上产生的嘈杂声和汽车鸣笛声，以及各种施工工地上产生的机器轰鸣声等。而利用计算机来自动提取这些声场景并对其进行分类具有重要的应用价值，如，场景声频监控[1]、设计助听器[2]、构建智能房间[3]和制造智能汽车等。

目前，对真实环境中的声场景即声事件进行精准的自动分类，还存在较大的困难。因为在真实的声场景中，通常会同时出现多种声事件，这导致某类声事件会受到其他背景声的干扰，从而使机器自动识别变得困难。因此，声场景分类具有重要的研究价值。近些年随着卷积神经网络 (Convolutional neural network, CNN) 的发展，出现了许多基于CNN的声场景分类方法，其中时频卷积神经网络 (Temporal-spectral convolutional neural network, TS-CNN) 提出了时频注意力模块[4]，是目前声场景分类效果最好的网络之一，但是由于其结构复杂且层数较多，导致其运算效率较低，推理开销大。为了提高性能，当前网络都是朝着更重、更复杂的方向发展，但是大型网络对搭载设备要求高，且运算速度慢，不利于实际应用。因此如何能够在不增加推理计算量的情况下提高网络的声场景分类能力，成为一大难题。

在不提高网络参数量的前提下，已有的提高深度卷积神经网络性能的方法包括协同学习 (Collaborative learning)[5]、多任务学习[6]和知识蒸馏[7]等。其中，协同学习是在网络的中间层连接额外的分类器对中间层进行直接监督。多任务学习是把多个相关任务放在一起学习，通过设计多个损失函数同时学习多个任务。而知识蒸馏是将已经训练好的大型教师网络中包含的知识，蒸馏提取到小型的学生网络。2015 年，Hinton 等[7]提出了知识蒸馏的方法，成功实现了网络与网络之间的知识转移，但是知识蒸馏方法存在多网络训练，且设计复杂的缺点。2016 年，Søgaard 等[8]证明了多任务学习的性能取决于多个相关任务的相似性，而在声场景分类中难以找到合适的相似任务。2018 年，Song 等[5]对协同学习中辅助分支的设计和不同引入中间层位置的选择进行了研究，研究证明简单的添加辅助分类器并不能提高网络的性能，而经过对辅助分支的结构进行设计和选择恰当的引入中间层位置可以有效提高网络性能。所以本文采用协同学习来对网络进行改进。

本文提出了一种基于协同学习的时频卷积神经网络模型 (TSCNN-CL)，能够在保持推理计算量不变的前提下，有效提高网络的声场景分类性能。本文的主要贡献包括：(1) 提出了在网络靠前的中间层上附加辅助监管分支，这些辅助监管分支可以起到一个鉴别中间层提取特征图的质量的作用。(2) 设计了一种同构分支结构，该结构可以提高主干网络的声场景分类性能。(3) 设计了一种基于 KL 散度的协同损失函数，在主干网络与辅助监管分支之间实现了成对知识交流，从而起到了正则化的作用，提高了网络的鲁棒性。(4) 采用了一种基于协同学习的测试策略，在测试时将辅助监管分支屏蔽，保持推理量不变，使模型便于工业部署中的实际应用。本文将所提出的模型在 ESC-50、ESC-10 和 UrbanSound8k 三个常用声音分类数据集上进行了实验验证，实验结果表明所提出的 TSCNN-CL 模型的平均分类准确率分别为 84.6%、93.5% 和 84.5%，相比于在 TS-CNN 模型上的实验结果分别提升了 1.2%、1.5% 和 1.0%。

1 声场景的特征提取

由于所需识别的声事件常常被背景噪声所掩盖，因此准确地提取其特征是声场景分类的关键。目前常用声音特征提取方法有短时傅里叶变换 (Short-time Fourier transform, STFT)、小波谱图和 Mel 谱图。其中，STFT 的方法是采用一个窗口函数，将声信号分割成许多小的时间间隔，然后对每一个时间间隔做傅里叶变换，以确定该时间间隔的频率；小波谱图是通过对声信号进行多尺度分解，将声信号分解到不同尺度上进行表示[9]，从而得到声信号的时频表达；而 Mel 谱图是基于人类听觉系统对不同频率尺度的感知，在 STFT 基础上进一步提取具有不同频率成分的特征信息，与 STFT 和小波变换相比，它提供更集中的声音频谱表示。由于这些时频表达方法得到的频谱图可以看成一幅图像，因此也可以采用图像处理的方法对其特征进行进一步描述，常用的方法如局部二进制模

式(Local binary patterns, LPB)或方向梯度直方图(Histogram of oriented gradient, HOG)等[10]。

上述声音特征提取方法只适合对特定领域的声信号进行表达。而对数梅尔谱图法(Log-Mel)通过对梅尔谱图取对数，压缩了频率的尺度，使特征变化更加平稳。同时避免了梅尔谱图因频率相差过高而导致的数据计算困难、低频率数据容易被忽视等问题，能够对不同领域的声信号进行更准确的表达。为此，本文选择Log-Mel谱图对声音特征进行表达。图1展示了一段烟火声的Log-Mel谱图。

2　时频卷积神经网络

时频卷积神经网络(TS-CNN)是由Wang等[4]提出的用于声场景分类的CNN，弥补了此前网络在提取深层特征时没有充分利用声音特有的频率和时间特征的缺陷。TS-CNN在CNN中引入时间—频率平行注意力机制，通过根据不同时间帧和频带的重要性进行加权对时间和频谱特征进行有选择的学习，同时平行分支构造可以分别应用时间注意力和频谱注意力，有效避免了噪声干扰。

TS-CNN的网络结构如图2所示。它由4个时频卷积模块(TFblock)组成，分别具有64、128、256和512个输出通道。其中每个卷积模块包含2个卷积层，卷积核大小为3 × 3，提取的对数梅尔谱图先通过时频注意力模块进行提取特征，然后经过平均池化层进行下采样，最后连接全局池化层和全连接层。在每个卷积层后都采用批量归一化层[11]和ReLU[12]激活函数。4个卷积层模块依次相连，使用Softmax分类器进行分类。

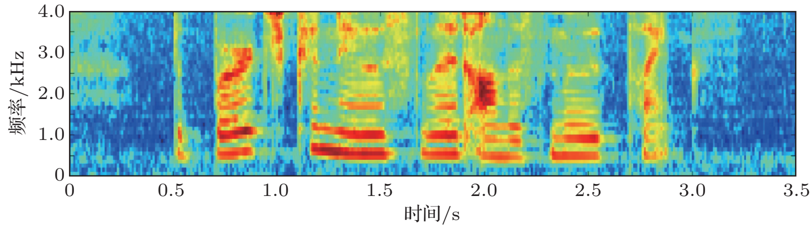


图1　烟火的对数梅尔谱图示例

Fig. 1　Example of Log-Mel of pyrotechnics

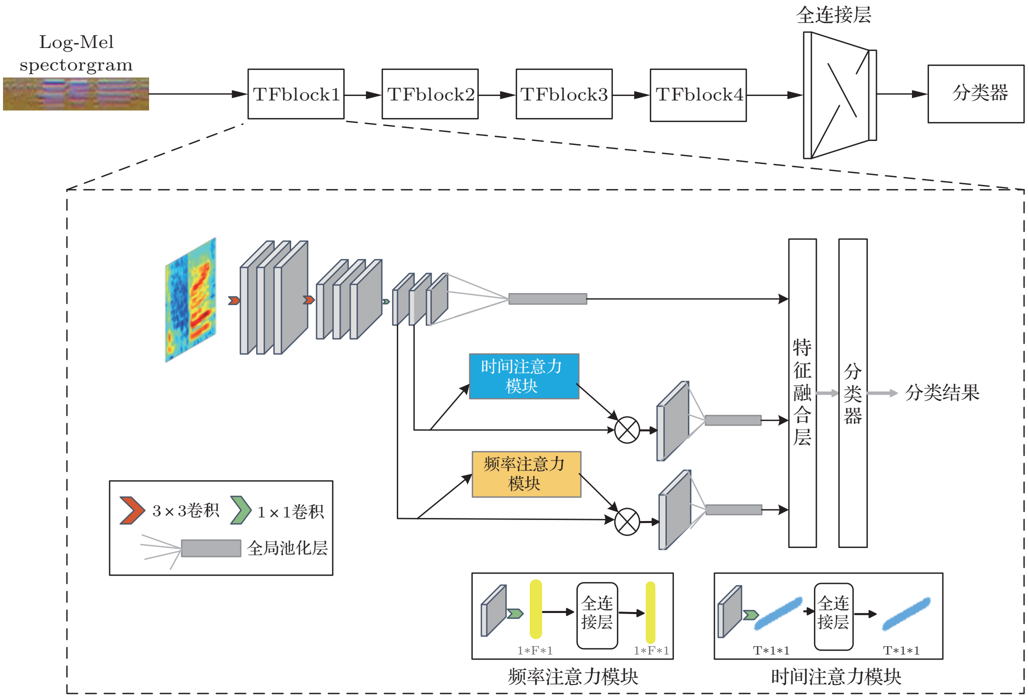


图2　TS-CNN结构框图

Fig. 2　TS-CNN model framework

TS-CNN 可充分利用声音固有的频率和时间特征，能够有效降低噪声的干扰，但由于 TS-CNN 网络层数较深，且在训练时采用非凸优化算法，导致网络在训练的时候，容易陷入局部最优值，并且伴随着梯度消失和梯度爆炸的现象，因此达不到最优效果。为了解决这一问题，在不增加推理量的前提下提高性能，本文在 TS-CNN 的基础上引入了协同学习，提出了 TSCNN-CL 网络。

3 协同时频卷积神经网络

协同时频卷积神经网络 (TSCNN-CL) 是在 TSCNN 基础上引入了协同学习的方法，通过增加两个协同分支以使得网络训练更加充分。增加 CNN 的深度虽然可以一定程度上提高网络的表征能力，但随着深度加深，会逐渐出现神经网络难以训练的情况，其中就包括像梯度消失和梯度爆炸等现象。为此，TSCNN-CL 在神经网络的中间层引入辅助的分支分类器，辅助分支分类器能够判别中间层提取的特征图质量的好坏，并且为中间层提供直接的监督，而不是 CNN 通常采用的仅在输出层提供监督，然后将此监督传播回早期层的标准方法。并且为每个分支设计了基于 KL 散度的辅助损失函数，使分支和主干之间进行信息交互，提高了网络的泛化能力。

3.1 网络结构

TSCNN-CL 的模型结构如图 3 所示。具体地，先将 TF 模块 1、TF 模块 2 和 TF 模块 3 的输出分别标记为 C、B、A 位，然后从 C 位和 B 位分别引出两条同构分支，在分支之间进行 KL 散度计算作为协同损失函数。其中，同构分支的网络结构与主干网络的网络结构完全相同。

3.2 协同损失函数

在 TSCNN-CL 中，两个协同分支采用交叉熵作为损失函数。而为了实现不同分类器之间的知识协同，在不同分支之间设计了一种基于 KL 散度的协同损失函数，使得连接到主干网络的所有分支之间可以进行信息交流，进一步优化网络性能。

设 $\mathbf{D} = \{(x_i, y_i|1 \leqslant i \leqslant N\}$ 为包含 N 个样本的数据集，其中 x_i 是第 i 个训练样本，y_i 是对应的真实标签。此外，设 $f(\boldsymbol{W}, x_i)$ 为 CNN 的输出向量。对于只在网络的最后一层增加监督的标准训练方案，优化目标可表示为

$$\underset{\boldsymbol{W}_1}{\operatorname{argmin}}\, L_1(\boldsymbol{W}_1, \mathbf{D}) + \lambda R(\boldsymbol{W}_1), \tag{1}$$

其中，L_1 为默认损失，R 为正则化项，λ 是正则化系数。在公式 (1) 中，L_1 由式 (2) 计算：

$$L_1(\boldsymbol{W}_1, \mathbf{D}) = \frac{1}{N}\sum_{i=1}^{N} H\left(y_i, f\left(\boldsymbol{W}_1, \boldsymbol{x}_i\right)\right), \tag{2}$$

其中，$H(\ )$ 是交叉熵损失函数，定义为

$$H\left(y_i, f\left(\boldsymbol{W}_1, x_i\right)\right) = -\sum_{k=1}^{K} y_i^k \lg f^k\left(\boldsymbol{W}_1, x_i\right). \tag{3}$$

对于 TSCNN-CL，因为分别在 B 位、C 位引出了协同分支，所以模型的优化目标为

$$\begin{aligned}\underset{\boldsymbol{W}_1, \boldsymbol{W}_\mathrm{B}, \boldsymbol{W}_\mathrm{C}}{\operatorname{argmin}}\quad & L_1(\boldsymbol{W}_1, \mathbf{D}) + L_\mathrm{B}(\boldsymbol{W}_\mathrm{B}, \mathbf{D}) + L_\mathrm{C}(\boldsymbol{W}_\mathrm{C}, \mathbf{D}) \\ & + L_\mathrm{AUX} + \lambda R(\boldsymbol{W}_1),\end{aligned} \tag{4}$$

其中，$\boldsymbol{W}_\mathrm{B}$、$\boldsymbol{W}_\mathrm{C}$ 分别为分支 B、C 的输出向量，L_AUX 为辅助损失函数。L_AUX 可表示为

$$\begin{aligned}L_\mathrm{AUX} = {} & \mathrm{KL}(\boldsymbol{W}_1 \mid \boldsymbol{W}_\mathrm{B}) + \mathrm{KL}(\boldsymbol{W}_\mathrm{B} \mid \boldsymbol{W}_1) \\ & + \mathrm{KL}(\boldsymbol{W}_1 \mid \boldsymbol{W}_\mathrm{C}) + \mathrm{KL}(\boldsymbol{W}_\mathrm{C} \mid \boldsymbol{W}_1) \\ & + \mathrm{KL}(\boldsymbol{W}_\mathrm{B} \mid \boldsymbol{W}_\mathrm{C}) + \mathrm{KL}(\boldsymbol{W}_\mathrm{C} \mid \boldsymbol{W}_\mathrm{B}).\end{aligned} \tag{5}$$

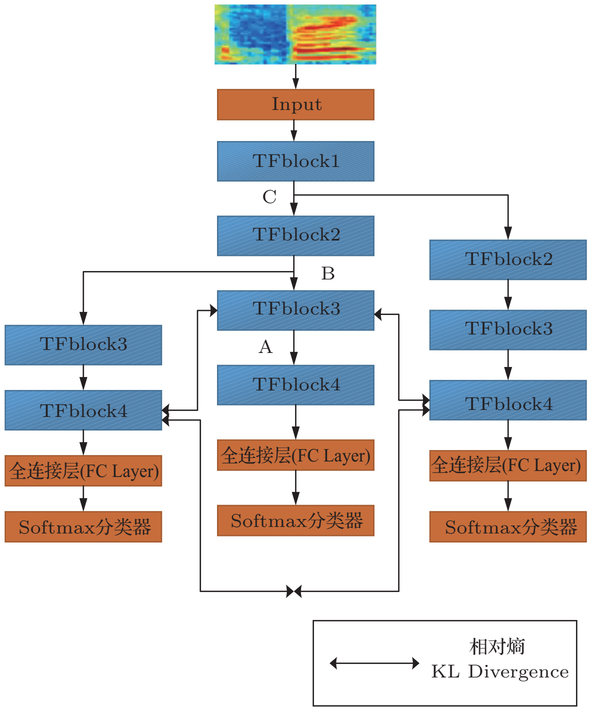


图 3 TSCNN-CL 模型结构图

Fig. 3 TSCNN-CL model framework

因为KL散度不具有交换性，TSCNN-CL的3条支路两两交互，因此设计了6个KL散度来组成辅助损失函数L_{AUX}。

4 实验结果与分析

为验证所提TSCNN-CL网络模型的有效性，本文在ESC-10、ESC-50和UrbanSound8k三个常用基准声音数据集上进行了分类实验验证。

4.1 数据库

(1) ESC-50/ESC-10[13]：ESC-50数据集是由2000个环境音频记录的集合，是一个适用于声场景分类的基准数据集。数据集中每个记录由5 s长的录音组成，分为50个小语义类(每个类有40个样本)。其中声频的采样频率为44.1 kHz。所有数据集被分为5个子集进行交叉验证，本文中将采用交叉验证结果的平均对网络性能进行评估。而ESC-10数据集是ESC-50数据集的一个子集，包含10个类别，每类40个例子。ESC-10数据集的所有其他特征都与ESC-50数据集相同。

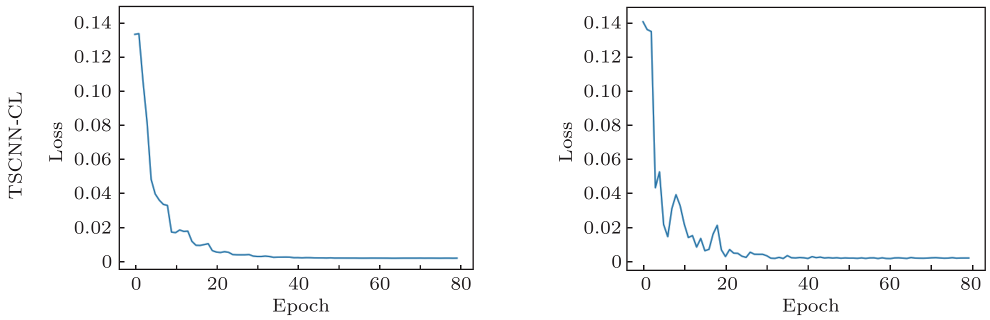
(2) UrbanSound8k[14]：Urbansound8k是目前应用最为广泛的公共数据集，主要用于自动城市环境声分类研究。UrbanSound8k数据集由8732个声频片段组成，一共分为10类：“空调”“汽车喇叭”“儿童玩耍”“狗叫”“钻孔”“发动机空转”“枪声”“风钻”“警笛”“街头音乐”。每个类的总声频时长是不均衡的，且每个声频样本的时长可变，最长是4 s，最短是2 s。样本采样频率从16 kHz到48 kHz不等。实验使用官方的10个交叉验证数据集进行模型性能评价。

4.2 数据预处理

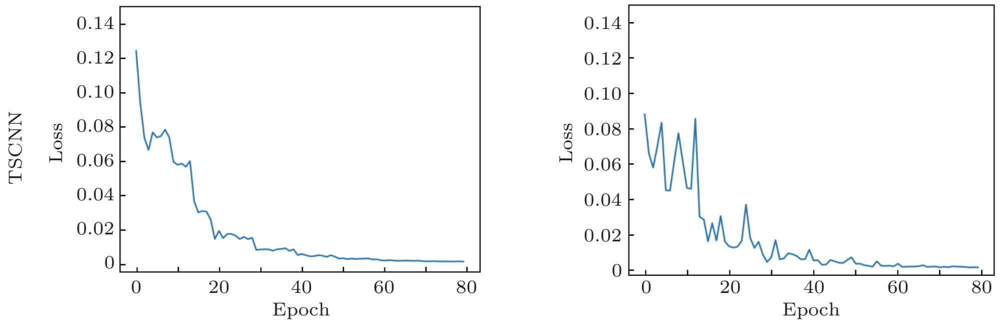
本文首先将所有的原始声频样本重新采至44.1 kHz，并且通过零填充将声频补充到同一长度：ESC-10和ESC-50扩充到5 s，UrbanSound8k扩充到4 s。然后采用STFT提取声频样本的谱图，设定的窗口大小为40 ms，跳跃大小为20 ms。最后通过梅尔滤波器得到对数梅尔频谱图。

4.3 网络训练

在进行网络训练时，本文选择Adam算法作为优化器，使用默认参数，初始学习率设置为0.03，指数衰减率为0.99。协同分支在训练时与主干网络一同训练，在推理时将其屏蔽，不增加额外推理代价。该网络由PyTorch实现，并且在Tesla V100上进行训练。图4为网络训练过程中的损失函数变化曲线。



(a) TSCNN-CL在训练过程中的损失函数曲线 (b) TSCNN-CL在验证过程中的损失函数曲线



(c) TSCNN在训练过程中的损失函数曲线 (d) TSCNN在验证过程中的损失函数曲线

图4 TSCNN-CL与TS-CNN的训练过程中损失函数变化曲线对比

Fig. 4 Comparison of loss changes in TSCNN-CL and TS-CNN models during the training process

由图4可以看出，在TSCNN-CL训练过程中，在迭代10 Epoch之前训练集和验证集的损失值从0.14迅速下降，在10 Epoch和30 Epoch之间损失函数缓慢下降，40 Epoch之后的损失值逐渐趋于平稳，且稳定在0.015。由于采用的验证集数据样本和训练集样本不同，两个模型在验证时损失值在20 Epoch左右存在震荡。此外，在与TSCNN的比较中可以看出，TSCNN-CL的损失函数曲线变化更加平滑，收敛更加迅速。

4.4 单分支与多分支比较

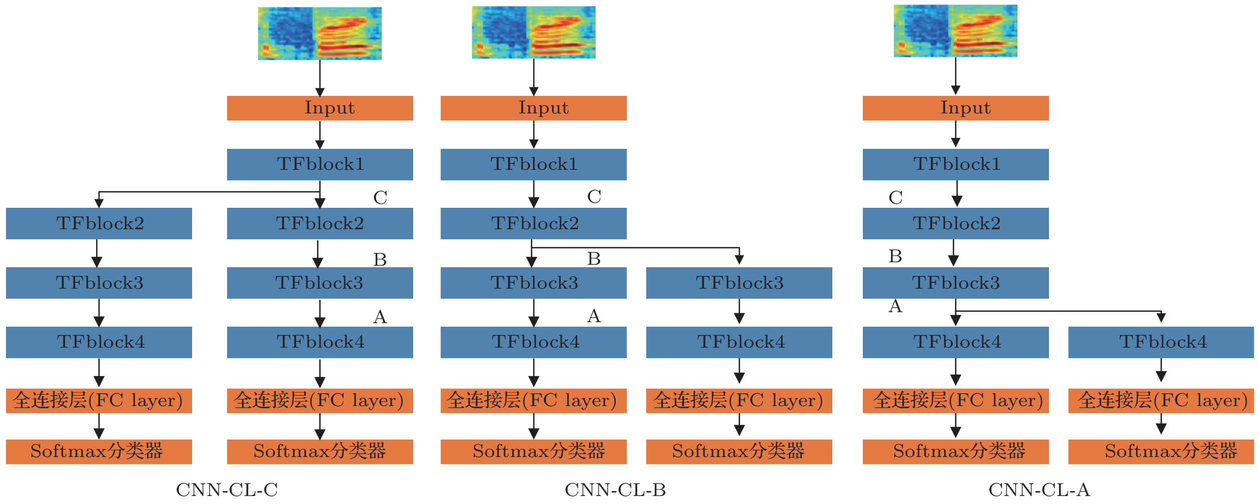
为验证多分支协同学习的有效性，本文分别在A位、B位和C位引出同构协同分支进行测试。图5分别展示了对应3个位点的网络结构。不同位点分支实验结果的分类正确率如表1所示。从表1可以看出，分支位点的位置越靠前，网络的性能越好。这是因为在网络的训练过程中随着迭代次数的增加，CNN早期层的卷积核参数的变化会趋于平缓。但这并不意味着早期层输出的特征图已经达到了最好的效果，而只是达到了一个局部最优。换言之，整体网络的性能由于早期层的卷积核没有得到充分的训练，而导致最终的分类效果没有得到提升。TSCNN-CL则通过对早期的卷积层添加协同分支，使其继续进行训练，从而提高了其输出的特征图质量，因此增强了网络的分类性能。

表1 不同分支之间的实验结果比较

Table 1 Comparison of experimental results among different branches

(单位：%)

Model	ESC-10	ESC-50	UrbanSound8k
TSCNN-CL-A	92.20	83.60	83.70
TSCNN-CL-B	92.30	83.70	83.80
TSCNN-CL-C	92.80	83.90	84.00
TSCNN-CL-BC	93.50	84.60	84.50



CNN-CL-C CNN-CL-B CNN-CL-A

图5 不同分支的框架

Fig. 5 The frameworks of different branches

4.5 实验结果比较与分析

为了验证TSCNN-CL模型的性能，本文将其与当前主流方法进行了比较。通过交叉验证，实验结果表明所提出的TSCNN-CL的平均分类准确率在ESC-50、ESC-10和UrbanSound8k上分别为84.6%、93.5%和84.5%，在TS-CNN实验结果的基础上分别提升了1.2%、1.5%和1.0%。其中TS-CNN的结果是按照作者给出的代码在相同实验环境下进行复现得到的。声场景分类的主流方法中，按照对声信号的与处理方式，可以分为两大类，分别是人工设计特征和原始声信号。人工设计特征是指声场分类任务从原始声信号中提取人工设计的特征，比如：时频图、梅尔图、梅尔倒谱系数作为神经网络的输入进行训练。2017年，谷歌将GoogLeNet [15]应用到了声场分类中，其采用梅尔图与梅尔倒谱系数相结合的方式对声信号进行预处理，取得了良好的分类效果。但在实际声场景中，声信号与语声和音乐信号不同，面临着录制条件复杂、噪声较多等问题，人工设计的特征无法对声信号的特征进行自适应的表示。而原始声信号方案可以利用神经网络

强大的特征提取能力，从声信号中提取出自适应的特征，同时也省去了复杂的人工设计特征过程。鉴于此优势，一些基于原始声信号的研究相继出现。2017 年，Tokozum 等[16] 提出了一种称为 EnvNet 的一维体系结构，它使用原始声信号作为输入进行端到端的训练，在当时达到了最好的分类效果。2019 年，Abdoli 等[15] 提出了 Gammatone 1D-CNN，模拟 Gammatone 滤波器组进行网络初始化，有效提高了网络的分类性能。尽管原始声信号方案与人工设计特征方案相比存在优势，但是由于一维的声信号比手工设计特征包含更多的噪声信息，并且神经网络需要大量的声音数据用于训练，而声音数据的获取难度要高于图像和文本数据，所以目前的主流方案还是人工设计特征方案。

表 2　TSCNN-CL 模型在 ESC-10、ESC-50 和 UrbanSound8k 上与其他声场景分类模型的对比

Table 2　Comparisons between TSCNN-CL model and other environmental sound classification models on ESC-10, ESC-50, and UrbanSound8k datasets

(单位: %)

Model	ESC-10	ESC-50	UrbanSound8k
Human[13]	95.70	81.30	
GoogLeNet[15]	86.00	73.00	93.00
Envnet[16]	88.10	74.10	71.10
Piczak-CNN[17]	90.20	64.50	73.70
Envnet v2[18]	91.30	84.70	78.30
VGG-like CNN[19]	91.70	83.90	83.70
GTSC+TEO-GTSC[20]		81.90	88.00
Gammatone 1D-CNN[21]			89.00
TS-CNN[4]	92.00	83.40	83.50
TSCNN-CL-BC(ours)	**93.50**	**84.60**	**84.50**

此外，GoogLeNet 在 UrbanSound8k 上的测试并没有按照标准划分 10 个子集进行交叉验证，而是采用了 5 个随机划分的交叉验证集。而 Gammatone 1D-CNN 虽然在 UrbanSound8k 分类效果较好，但主要是对声音特征进行了重叠提取，提取的相邻特征信息之间存在 50% 的重叠，相当于对数据进行了增强，且测试集里包含了训练集的样本，因而提升了分类效果。TSCNN-CL 与其他主流方法相比，采用了时频注意力模块对声信号的时间和频率特征进行加权学习，不仅能够有效避免噪声的干扰，而且通过引入协同学习，能最大程度地挖掘网络潜力，进一步增强了网络的分类性能。表 2 显示了 TSCNN-CL 和其他主流方法的性能比较，结果表明，本文提出的协同学习的方法能够显著提高网络的分类效果。

5　结论与展望

本文提出了一种基于协同学习的时频卷积神经网络 (TSCNN-CL) 用于声场景自动分类。TSCNN-CL 通过协同学习的方法，在不增加推理量的前提下，提高了网络的分类性能。首先在 TS-CNN 的中间层引入两条协同分支，这两条协同分支能够辅助监督中间层训练。其次在主干与分支之间设计了相应的辅助损失函数，使得主干和分支可以进行信息交互，提高了网络的泛化能力，并且为协同分支之间也设计了协同损失函数，实现了分支之间的成对知识匹配。最后，在推理的时候将分支屏蔽，保持推理运算量不变，使模型便于工业部署。在声场识别常用数据集 ESC-10、ESC-50 和 UrbanSound8k 上的实验结果表明所提出的 TSCNN-CL 网络模型的分类效果较 TS-CNN 模型有较大提升，且优于当前大部分的主流方法。

参考文献

[1] Radhakrishnan R, Divakaran A, Smaragdis A. Audio analysis for surveillance applications[C]//IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005: 158–161.

[2] Giannoulis D, Stowell D, Plumbley M. Acoustic scene classification: classifying environments from the sounds they produce[C]// IEEE Signal Processing Magazine, 2015: 16–34.

[3] Vacher M, Serignat J F, Chaillol S. Sound classification in a smart room environment: an approach using GMM and HMM methods[C]// The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue, 2007: 135–146.

[4] Wang H, Zou Y, Chong D, et al. Environmental sound classification with parallel temporal-spectral attention[C]// INTERSPEECH 2020, 2020: 25–29.

[5] SongG, Chai W. Collaborative learning for deep neural networks[C]// the 32nd International Conference on Neural Information Processing Systems (NIPS'18), 2018: 1832–1841.

[6] Caruana R A. Multitask learning: a knowledge-based source of inductive bias[C]// In Proceedings of the Tenth

International Conference on Machine Learning, 1993: 41–48.

[7] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, arXiv: 1503.02531v1.

[8] Søgaard A, Goldberg Y. Deep multi-task learning with low level tasks supervised at lower layers[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 231–235.

[9] Zhao R, Pandit V, Qian K, et al. Deep sequential image features on acoustic scene classification[C]// Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[10] Chen H, Liu Z, Liu Z, et al. Integrating the data augmentation scheme with various classifiers for acoustic scene modeling[J]. arXiv Preprint, arXiv: 1907.06639.

[11] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[J]. arXiv Preprint, arXiv: 1502.03167, 2015.

[12] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning, 2010: 807–814.

[13] Piczak K J. ESC: dataset for environmental sound classification[C]//Proceedings of the 23rd Annual ACM Conference on Multimedia, 2015: 1015–1018.

[14] Salamon J, Jacoby C, Bello J. A dataset and taxonomy for urban sound research[C]// Proceedings of the 22nd ACM international conference on Multimedia, 2014: 1041–1044.

[15] Boddapati V, Petef A, Rasmusson J, et al. Classifying environmental sounds using image recognition networks[C]// Procedia Computer Science, 2017: 2048–2056.

[16] Tokozume Y, Harada T. Learning environmental sounds with end-to-end convolutional neural network[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: 2721–2725.

[17] Piczak K J. Environmental sound classification with convolutional neural networks[C]//IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015: 1–6.

[18] Tokozume Y, Ushiku Y, Harada T. Learning from between-class examples for deep sound recognition[C]// arXiv Preprint, arXiv: 1711.10282.

[19] Zhang Z, Xu S, Cao S, et al. Deep convolutional neural network with mixup for environmental sound classification[C]//Pattern Recognition and Computer Vision (PRCV), 2018: 356–367.

[20] Agrawal D, Sailor H, Soni M, et al. Novel TEO-based Gammatone features for environmental sound classification[C]// in 2017 25th European Signal Processing Conference (EUSIPCO). 2017: 1809–1813.

[21] Abdoli S, Cardinal P, Koerich A. End-to-end environmental sound classification using a 1d convolutional neural network[C]//Expert Systems with Applications, 2019: 252–263.