王汝旭, 王荣燕, 曾科,等. 基于 Vision Transformer 和迁移学习的家庭领域哭声识别[J]. 智能计算机与应用,2024,14(6):119-126. DOI:10.20169/j.issn.2095-2163.240616

基于 Vision Transformer 和迁移学习的家庭领域哭声识别

王汝旭, 王荣燕, 曾　科, 杨传德, 刘　超
(德州学院 计算机与信息学院, 山东 德州 253023)

摘　要: 针对 SVM 等传统机器学习算法准确率低和当前使用 CNN 处理家庭领域哭声识别在不同婴儿间出现泛化能力差的问题,提出了一种基于 Vision Transformer 和迁移学习的婴儿哭声音频分类算法。首先,为实现数据集样本的扩增,采用了包括梅尔频谱转换和数据增强的数据预处理技术,进而达到了增强模型鲁棒性的目的。而后,在微调后的 Vision Transformer 模型上进行迁移学习训练,同时,训练过程中利用了 LookAhead 优化器来不断调整模型参数以避免过拟合,最终实验实现了对婴儿哭声音频的自动分类。实验结果表明,本实验模型相比其他深度学习模型具有更高的精确率和更快的收敛速度,同时还能有效地学习到婴儿哭声中更具区分性的特征。可以在新生儿监护、听力筛查和异常检测等领域中发挥重要作用。

关键词: Vision Transformer 模型; 婴儿哭声; 迁移学习; 梅尔频谱图; LookAhead
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 2095-2163(2024)06-0119-08

Classification of infant cry sounds based on Vision Transformer and transfer learning

WANG Ruxu, WANG Rongyan, ZENG Ke, YANG Chuande, LIU Chao
(School of Computer and Information, Dezhou University, Dezhou 253023, Shandong, China)

Abstract: Aiming at the low accuracy of traditional machine learning algorithms such as SVM and the poor generalization ability of the current CNN in dealing with cry recognition in the family field between different infants, the infant cry audio classification algorithm based on Vision Transformer and transfer learning is proposed. Firstly, in order to realize the expansion of the data set samples, the data preprocessing technology including MEL spectrum conversion and data augmentation is used, so as to achieve the purpose of enhancing the robustness of the model. Then, transfer learning training is performed on the fine-tuned Vision Transformer model. At the same time, the LookAhead optimizer is used to continuously adjust the model parameters in the training process to avoid overfitting. Finally, the research realizes the automatic classification of infant crying audio. The experimental results show that the proposed model has higher accuracy and faster convergence speed than other deep learning models, and can effectively learn more discriminative features in infant crying. The research can play an important role in the fields of neonatal monitoring, hearing screening and anomaly detection.

Key words: Vision Transformer networks; infant cry sounds; transfer learning; Mel spectrogram; LookAhead

0　引　言

随着科技的不断发展,人们对于保护婴儿的健康和安全越来越关注。在婴儿表达需求与不适时,哭声是一种主要的沟通方式。通常,婴儿哭声识别的方法包括对其信号进行端点检测和预处理,提取特征参数序列,并通过匹配算法来确定是否与已有模板序列相匹配。

文献[1]和文献[2]都探究了婴儿哭声识别的方案,并在理想情况下均获得了较高的正确率。其中,文献[1]使用 MFCC 参数作为特征提取方式,并采用 DTW 匹配算法进行哭声识别;而文献[2]则利用线性预测系数(LPC)作为特征提取方式,并同样运用 DTW 匹配算法进行分类。但是这类方法对于

基金项目: 国家级大学生创新训练项目(202210448014)。
作者简介: 王荣燕(1982-),女,博士,讲师,主要研究方向:模式识别,音频场景识别,音频事件分类等; 曾　科(2004-),男,本科生,主要研究方向:深度学习; 杨传德(2003-),男,本科生,主要研究方向:深度学习; 刘　超(2002-),男,本科生,主要研究方向:图像处理,图像识别。
通讯作者: 王汝旭(2003-),男,本科生,主要研究方向:模式识别,音频事件分类。Email:2570789775@qq.com
收稿日期: 2023-05-08

信号长度的变化和背景噪声的影响比较敏感,缺乏自适应学习和泛化能力,需要手动选择特定的特征和参数。因此,如何设计一种时间效率高、同时能够处理在不同婴儿个体之间存在信号长度变化的情况,实现准确识别并具有良好鲁棒性的婴儿哭声识别算法,已成为哭声识别领域的一个新问题。

针对上述问题,本实验采用了梅尔频谱图作为输入数据,梅尔频谱图是一种常用的音频表示方法,通过对音频信号进行傅里叶变换和滤波处理,将音频信号转化为在梅尔频率上的能量分布,从而提取出音频的特征信息。在模型设计方面,本实验使用了一种新的模型—Vision Transformer(ViT)[3],是一种完全基于注意力机制的图像分类模型,并且已经在其他图像分类任务上取得了优异的性能表现。因此,将之作为本实验中迁移学习策略的预训练的模型,这可以有效提高模型的泛化能力,加快训练速度并提高分类精度,为婴儿哭声在实际复杂场景下的分类提供新方案。

1 方法介绍

传统的分类方法主要依赖于手动提取特征和分类器的构建,但这种方法往往受到特征提取的限制,未能取得令人满意的分类效果。近年来,深度学习技术的发展为音频分类研发任务处理提供了新的可能。

目前在音频分类领域,卷积神经网络(CNN)和循环神经网络(RNN)是主流的模型,但是这些模型都需要大量的计算资源和训练时间。随着 ViT 模型的提出,该模型在图像领域的表现已经超过了传统的卷积神经网络,并且可以处理变长序列数据。因此,本文将探索 ViT 在音频分类任务中的应用潜力。同时,由于训练一个深度神经网络需要大量的数据和计算资源,迁移学习成为提高模型性能的常用策略之一。本文基于迁移学习的思想,在预训练的 ViT 模型上进行微调,并采用预处理技术来减少数据噪声和增加数据样本的多样性,提高模型的泛化能力。

1.1 模型搭建

1.1.1 Vision Transformer 网络

Vision Transformer(ViT)是一种新型的图像分类模型,与传统的卷积神经网络不同的是,ViT 采用了 Transformer[4-5] 中的自注意力机制。在 ViT 中,输入的是一组图像块(Patch),而不是整张图像,这些图像块被转换成向量表示并通过多层 Transformer 编码器(Encoder)进行处理。ViT 的输出是一个表示整张图像的向量,可以被用于分类任务。与传统的卷积神经网络相比,ViT 能够更好地捕捉图像中的特征,具有更好的性能。Vision Transformar 模型设计如图 1 所示。由图 1 可知,模型由 3 个模块组成,分别是:Linear Projection of Flattened Patches(Embedding 层)、Transformer Encoder 和 MLP Head。

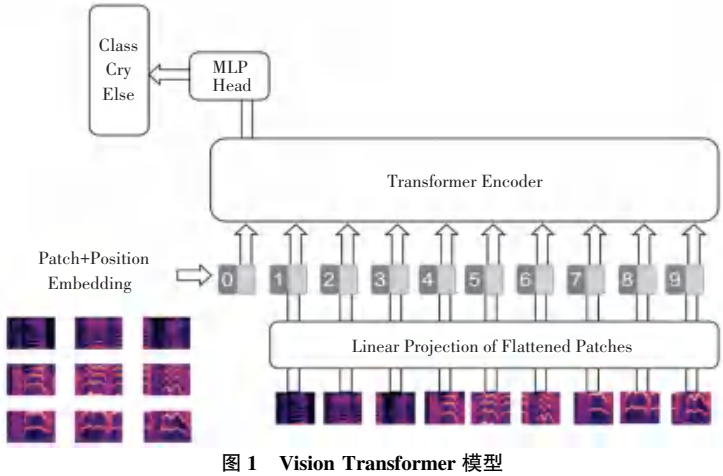


图 1 Vision Transformer 模型

Fig. 1 Vision Transformer model

1.1.2 Embedding 层

对于图像数据而言,其数据格式为 $[H, W, C]$ 的三维矩阵、而这却不是 Transformer 想要的。所以就要通过一个 Embedding 层来对数据进行变换。Embedding 层的设计如图 2 所示。图 2 中,首先将一张图片按给定大小分成一堆 Patches。以 ViT-B/16 为例,将输入图片(224×224)按照 16×16 大小的 Patch 进行划分,划分后会得到 $(224/16)^2 = 196$ 个 Patches。接着通过线性映射将每个 Patch 映射到一维向量中,以 ViT-B/16 为例,每个 Patche 数据结构为[16, 16, 3],通过映射得到一个长度为 768 的向量。

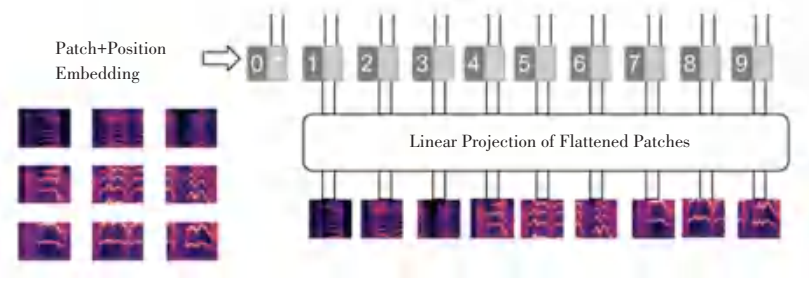


图 2　Embedding 层

Fig. 2　Embedding layer

1.1.3　Transformer Encoder

Transformer Encoder 是重复堆叠 Encoder Block L 次，如图 3 所示。由图 3 可看到，Encoder Block 的组成可做阐释分述如下。

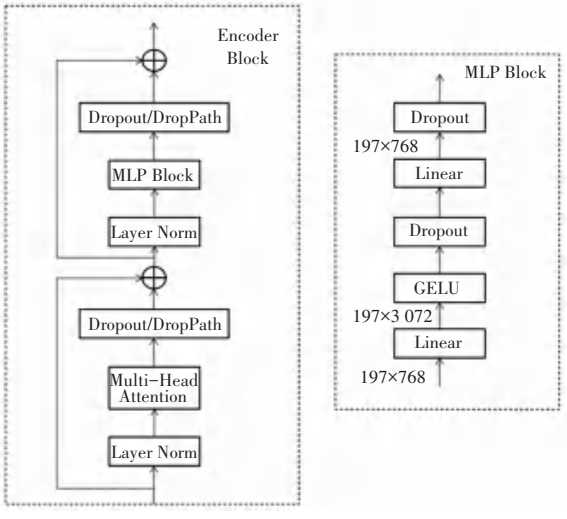


图 3　Transformer Encoder

Fig. 3　Transformer Encoder

(1) Layer Norm。是指对每个样本的所有特征在其维度上进行标准化，使得每个特征的均值为 0，方差为 1。这有助于网络在训练时更加稳定，因为不同样本的特征分布不同，标准化可以使得每个样本的特征都有相同的重要性，从而提高模型的泛化能力；

(2) Multi-Head Attention。Multi-Head Attention 是一种机制，用于处理输入数据。能使用多个不同的线性变换来映射输入数据到多个不同的向量空间中。然后，针对每个向量空间，计算一组注意力权重，以捕捉输入数据的不同方面之间的相关性。最后，将这些注意力加权的向量拼接起来，作为最终的输出表示。这种机制可以帮助模型更好地理解输入数据的多个方面，并捕捉数据间的复杂关系；

(3) Dropout/DropPath。Dropout 和 DropPath 是 2 种防止过拟合的技术，都可以通过随机断开神经元的连接来防止网络过度拟合训练数据。其中，Dropout 是指在训练时随机删除一部分神经元，而 DropPath 则是在训练时随机删除一部分连接；

(4) MLP Block。是指由 2 个全连接层组成的模块，其中第一个全连接层的输出通过一个非线性激活函数进行处理，再传递给第二个全连接层。

1.1.4　MLP Head

MLP Head 是指 Vision Transformer 模型中的最后一层，设计结构如图 4 所示。该层的输入为 Transformer Encoder 的输出特征向量，其维度通常较高，为了减少计算复杂度和过拟合的风险，通常会对其进行一些降维操作。降维完成后，特征向量会被送入 MLP 中，经过一层或多层全连接的神经元，每个神经元都会对输入特征进行加权、非线性变换和偏置操作，从而将特征向量转化为具有不同类别的概率分布或连续值的预测结果。

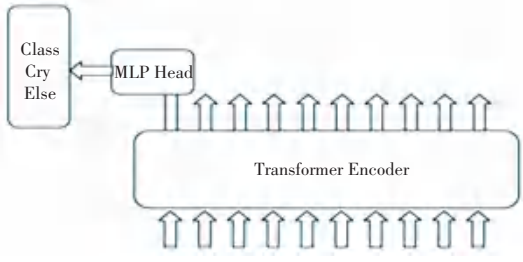


图 4　MLP Head

Fig. 4　MLP Head

1.2　梅尔频谱图

机器学习的第一步都是要提取出相应的特征(Feature)，如果输入数据是图片，例如 28 * 28 的图片，那么只需要把每个像素(Pixel)作为特征，对应的像素值大小(代表颜色的强度)作为特征值即可。在音频、语音信号处理领域中，语谱图(Spectrogram)是一种将信号在时域和频域上进行可视化的方法。具体是将信号在时间轴上分段，每段信号进行傅里叶变换得到对应的频域信息，然后将频域信息以颜

色或灰度值的形式表示出来,构成一个二维矩阵。这个二维矩阵中,横轴表示时间,纵轴表示频率,而矩阵中的每个元素则表示在对应的时间和频率上的信号强度,从而提供了一种直观的方式来展示信号的频率、时间和能量信息。

人类的听觉系统对声音频率的感知是基于对数刻度的模型。因此,语谱图中使用对数刻度来表示频率,以更好地模拟人类听觉系统的感知特性。这种方式使得研究时对低频的变化更加敏感,对高频的变化则相对迟钝。

研究表明,人类对频率的感知并不是线性的,并且对低频信号的感知要比高频信号更敏感。例如,人们可以比较容易地发现 500 Hz 和1 000 Hz的区别,却很难发现7 500 Hz 和8 000 Hz的区别。这时,梅尔标度(the Mel Scale)被提出,这是赫兹的非线性变换,对于以 Mel scale 为单位的信号,可以做到使得人们对于相同频率差别的信号的感知能力几乎相同。

所以线性分布的语谱图显然在特征提取上会出现“特征不够有用的情况”,因此梅尔频谱图应运而生。梅尔频谱图[6]的纵轴频率和原频率可由式(1)、式(2)来做互换:

$$m = 2\ 595\log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

$$f = 700\left(100^{m/2\ 595} - 1\right) \tag{2}$$

梅尔频谱图(Mel-spectrogram)是一种用于音频信号分析的可视化表示方法,将音频信号转换为时频域的矩阵表示。Mel-spectrogram 提取流程如图 5 所示。梅尔频谱图的获得方法具体如下。

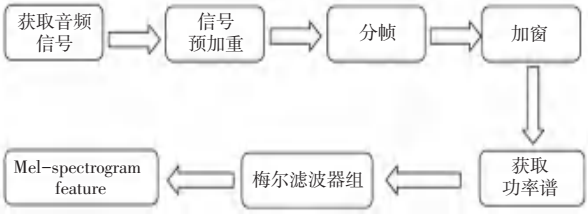


图 5 Mel-spectrogram 提取流程

Fig. 5 Mel-spectrogram extraction process

1.3 深度迁移学习

传统深度学习侧重在优化原有结构,专注于神经网络结构上的突破。迁移学习更趋向于在已有网络结构的基础上的领域适应性问题的研究。通常将源域(Source Domain)与目标域(Target Domain)分别表示为 D 和 T;$\mathcal{X}$ 与 y 分别表示特征空间与标签空间;$P(X)$ 表示边际概率分布,其中 $X = \{x_1, k, x_n\} \in \mathcal{X}$;$f(\cdot)$ 表示目标预测函数。迁移学习即借助有标注的 D 来学习 T 的知识,等同于如下形式:

$$D = \{\mathcal{X},\ P(X)\} \tag{3}$$

$$T = \{y,\ f(\cdot)\} \tag{4}$$

深度迁移学习的基本思想将一个已经在大规模数据集上训练过的深度神经网络模型(源模型),应用于另一个相关任务的学习过程。这个过程中,源模型的一些或全部层被保留下来,并被用于初始化目标任务的新模型。这种方法有助于解决许多实际问题,尤其是当新任务的数据集非常小的时候。

使用深度迁移学习,目标任务可以从源模型中学习到高级别的特征表示,这些特征通常是在源数据集上学习到的。通过保留这些层,新模型可以更快地收敛,并且可能达到更好的性能。这种方法通常被应用于图像识别、自然语言处理、语音识别等任务中,因为这些任务需要对大量的数据进行训练,并且深度学习模型的复杂性很高。

本文的实验模型在 ImageNet[7]这样一个超大型的数据集(包括约 120 万自然图像和1 000个不同类别)上进行预训练,将预训练得到的模型参数作为网络的初始化参数,此后依据经验定制全连接层,以此来增大感受野,进而学习到有婴儿哭声梅尔频谱图像更高细粒度特征模板。

1.4 LookAhead 优化器

LookAhead 优化器[8]是由 Zhang 等学者提出的,其目的是通过加强优化器的全局搜索能力来提高模型的泛化性能。这是一种新型的优化器,结合了快速和慢速两个优化器,并利用了预测梯度方向的思想,可以在一定程度上提高模型训练的速度和稳定性。LookAhead 优化器的核心思想是,引入一个“慢速”优化器,使其能够更好地探索搜索空间,并保持对目标的预测方向。具体来说,LookAhead 优化器包含 2 个部分:

(1)快速优化器。即原始的优化器,例如 Adam 或 SGD 等。本实验所用的是 Adam;

(2)慢速优化器。通过对快速优化器的参数进行移动平均得到,相对于快速优化器而言移动速度较慢。

LookAhead 优化器在每一次迭代中,先使用快速优化器进行梯度下降更新,并使用慢速优化器来探索一个更广的搜索空间。在下一次迭代中,将快速优化器的参数与慢速优化器的参数进行结合,得到一个更好的参数估计。使用 LookAhead 优化器可以加速模型的训练过程,并且可以提高模型的鲁棒性,避免在过拟合时出现局部最优解。

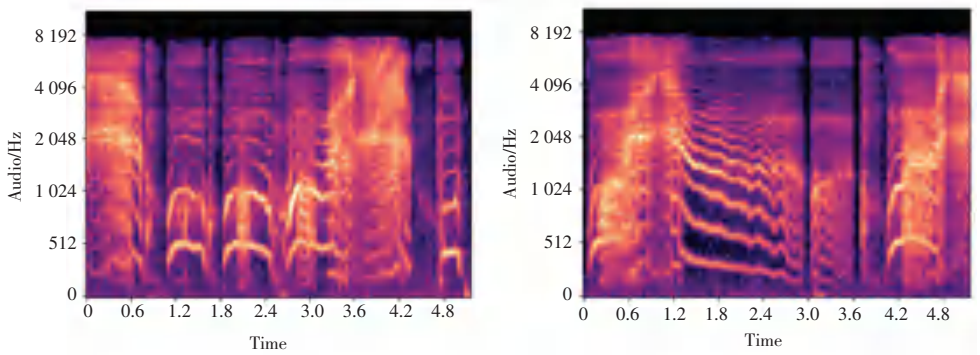
2　实验

2.1　实验环境

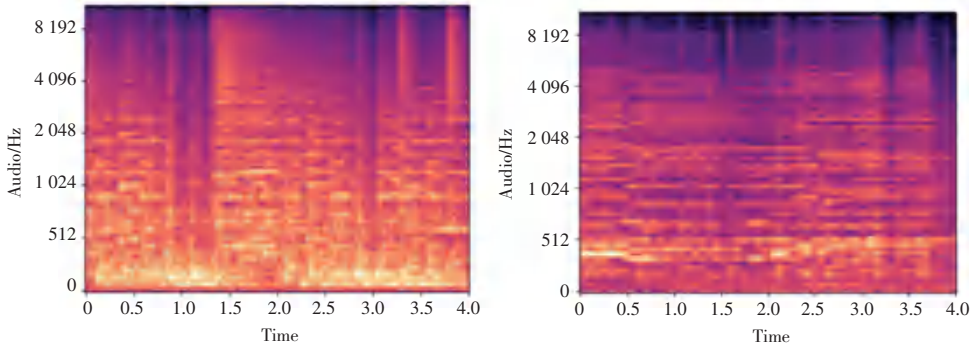
本实验采用 Python 编程语言，GPU 处理器为 NVIDIA GeForce GTX 3080Ti，内存为 16 GB，操作系统为 Ubuntu20.04，深度学习框架为 Pytorch。

2.2　实验数据

数据集婴儿啼哭声样本来自 PaddlePaddle 上的啼哭声数据集。数据集来源见表 1。数据集中，有的音频时长不相等，通过脚本将音频时长裁剪至 3 s 左右。通过裁剪后的音频得到 3 402 条婴儿哭声音频数据，然后采用 Urbansound8k[9] 数据集中大约 3 s 的 3 429 条非婴儿哭声音频数据。进一步将音频转换成梅尔频谱图，如图 6 所示。



(a) 哭声



(b) 非哭声

图 6　婴儿哭声和非哭声图像样本

Fig. 6　Samples of infant crying and non-crying images

表 1　数据集来源

Table 1　Data set sources

Name	Audio duration/s	Class	Audio quantity
Infant crying	≤15	awake	160
Infant crying	≤15	diaper	134
Infant crying	≤25	hug	160
Infant crying	≤25	hungry	160
Infant crying	≤20	sleep	144
Infant crying	≤25	uncomfortable	160
Urbansound8k	≤4	Air conditioner	1 000
Urbansound8k	≤4	Car whistle	429
Urbansound8k	≤4	Bark	1000
Urbansound8k	≤4	Children Playing	1 000

2.3　数据增强

音频数据增强是一种通过在训练数据上应用各种变换来生成新的数据样本，以增加训练数据的数量和多样性，从而提高深度学习模型的泛化能力的方法。数据集未增强和增强图像样本效果如图 7 所示。由图 7 可知，常用的音频数据增强技术包括：

(1)时间伸缩。改变音频信号的时长，可以通过线性插值或变换谱来实现。例如，将一个音频信号的时长拉长或缩短一定比例，可以生成新的数据样本。此实验将速度随机设置在 0.1~1.3 之间；

(2)噪声添加。向原始音频信号中添加白噪声、脉冲噪声、高斯噪声等，以模拟真实场景中的噪声干扰。例如，通过在婴儿哭声音频中添加噪声，可

以增加模型对于噪声干扰的鲁棒性。此实验将高斯噪声的标准差设置为 0.05；

(3)音调变换。音频数据增强中的音调变换(Pitch shift)是一种调整音频信号中所有频率的高低程度的技术。该技术可以用来模拟声音的高低变化,例如婴儿的声音在不同情绪下可能会有高低不同的音调。此实验将音调随机设置在-3~3 之间。

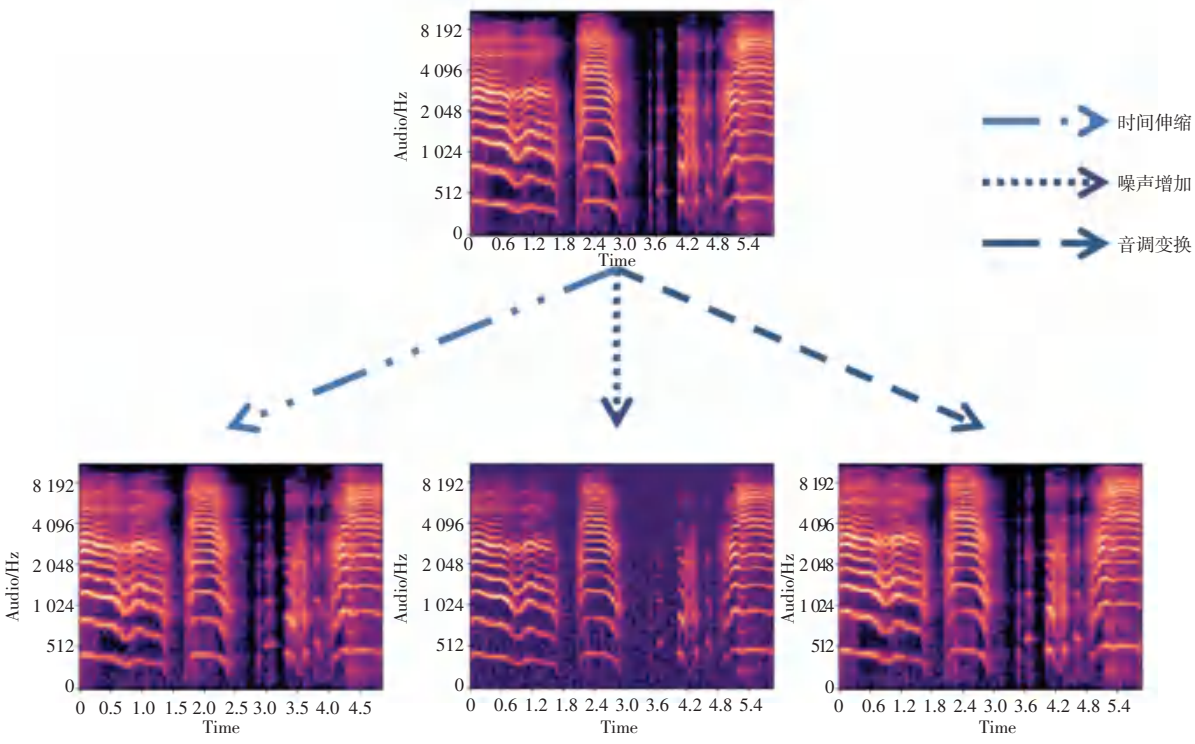


图 7 数据集未增强和增强图像样本

Fig. 7 Unenhanced and enhanced image samples of the dataset

通过以上的数据增强技术,可以生成更多且多样化的训练数据,从而提高深度学习模型的泛化能力和性能。同时,需要根据具体的任务场景和数据特征,选取适合的数据增强技术来提升模型的表现。

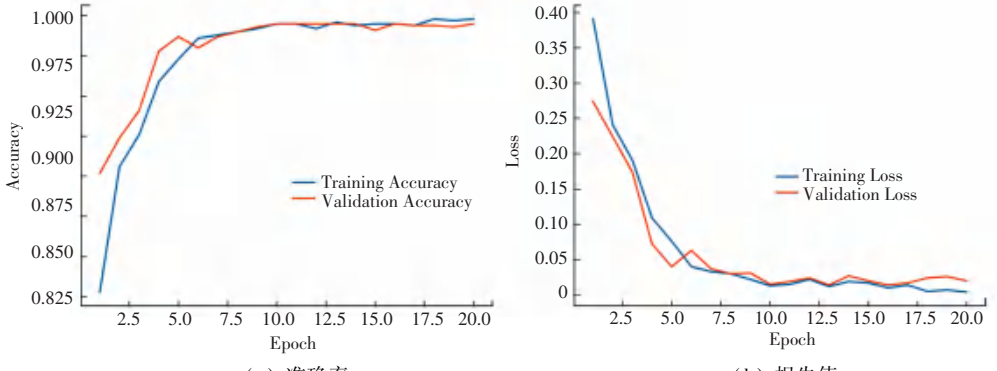
2.4 参数设置

选用 Vision Transformer 网络结构可以使得模型的精度和效率达到最优效果,同时也不会过度消耗计算机资源。经多次实验后,得到了本模型的最优参数组合,设置图像标准化 *Norm_size* 为 224,批量大小 *Batch_size* 为 16, 初始学习率为 0.000 1,共迭代了 epoch 20 次。

2.5 实验结果及分析

2.5.1 实验结果

根据常用性能评价指标,准确率越高,模型识别越准确,越能有效避免婴儿哭声未被识别出的情况;损失越低,则说明模型鲁棒性的强大。模型训练过程中准确率和损失变化情况如图 8 所示。实验结果表明:本实验最终验证集准确率和损失率分别为0.987 3和0.042 3。训练过程速度快,网络收敛性良好。



(a) 准确率 (b) 损失值

图 8 模型训练过程中准确率和损失变化情况

Fig. 8 Changes in accuracy and loss during model training

2.5.2　结果分析

数据集中训练集、验证集和测试集已被分好，未进行预处理前，训练集的数据量为 5 464 张，验证集的数据量为 1 366 张，测试集数据量为 800 张。为了更准确地验看识别效果，对测试集上的图像进行预测，并用混淆矩阵查看预测结果。

实验模型在测试集上的混淆矩阵如图 9 所示。从图 9 中可以看出，在测试集上，对于哭声和非哭声梅尔谱图图像，准确率均达到 98%左右，*recall* 达到 0.990 ，*f*1 达到0.987 5，可见本模型对于哭声梅尔频谱图图像分类效果理想。

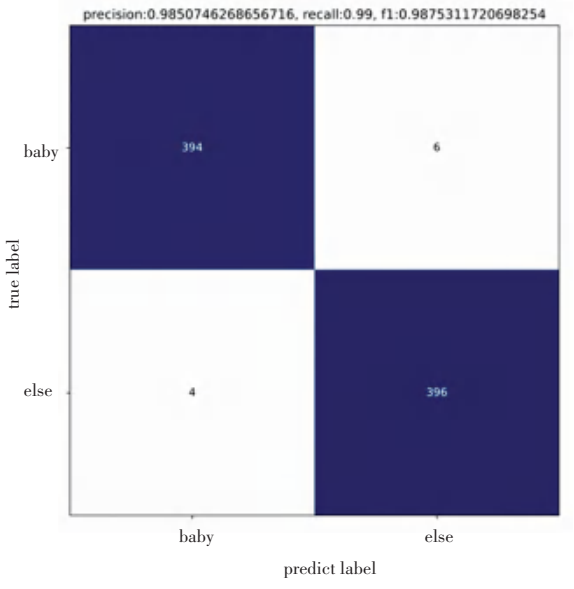


图 9　实验模型在测试集上的混淆矩阵

Fig. 9　Confusion matrix of experimental model on test set

在模型训练过程中，可以看出模型相较于轻量级神经网络模型 EfficientNet[10]，准确率更高；相较于其他深度学习网络模型如 ResNet[11]、VGG16[12]，准确率较高。对比数据见表 2。

表 2　数据集上深度学习分类模型的性能指标

Table 2　Performance indicators of the deep learning classification model in the data set

模型	输入特征	*Accuracy*	*Loss*	*Param*/ MB
EfficientNetV2-B2	Mel Spectrogram	0.926 1	0.332 8	32.0
VGG16	Mel Spectrogram	0.954 2	0.147 2	59.5
ResNetV2-50	Mel Spectrogram	0.969 2	0.063 7	94.8
Vision Transformer	Mel Spectrogram	0.985 0	0.042 3	85.3
EfficientNetV2-B2	MFCC	0.903 3	0.273 6	31.6
VGG16	MFCC	0.954 4	0.057 0	58.9
ResNetV2-50	MFCC	0.967 9	0.060 1	94.6

本实验还将模型与传统的分类模型做实验对比、如 SVM[13]，相较于传统分类模型有着更高的准确率。对比数据见表 3。

表 3　数据集上传统分类模型的性能指标

Table 3　Performance indicators of traditional classification models on the data set

模型	输入特征	*Accuracy*
SVM	Mel Spectrogram	0.914 6
SVM	MFCC	0.907 5
Vision Transformer	Mel Spectrogram	0.985 0

此外，本实验通过与其他优化器（如 SGD[14]、Adam[15]）的对比实验来评估 LookAhead 优化器的性能，在相同的训练周期，使用不同的优化器选择合适的学习率并记录模型的最佳测试精度。对比数据见表 4。

表 4　优化器在本模型的性能指标

Table 4　Performance indicators of the optimizer in this model

模型	Optimizer	Learning rate	*Accuracy*	*Epoch*
Vision Transformer	SGD	0.000 1	0.880 7	20
Vision Transformer	Adam	0.000 1	0.978 9	20
Vision Transformer	LookAhead	0.000 1	0.985 0	20

观察以上表中数据，可以明显看出，基于 Vision Transformer 和迁移学习的婴儿哭声音频分类算法在参数量和准确率上的优越性。

3　结束语

随着人工智能技术的不断发展，音频分类已经成为了一个热门的研究领域。婴儿哭声是婴儿沟通的一种方式，同时也是其表达需求和传达状态的重要方式。因此，基于婴儿哭声的识别能够辅助父母更好地了解和照顾婴儿。本文使用了一个经过预训练的 Vision Transformer 模型，即 ViT-B/16[3]作为基础模型，并通过在婴儿哭声数据集上进行微调，以实现对婴儿哭声的有效分类。同时，本文还采用了迁移学习的方法，即将在其他任务中训练好的模型参数应用到本任务中，以加速模型的训练和提高准确率。本实验采用了开源的婴儿哭声数据集，并将其划分为训练集和测试集。通过实验结果表明，本实验的模型在测试集上达到了 98%左右的分类准确率，优于其他常用的分类算法，例如卷积神经网络和 SVM 等。在优化算法方面，本文采用了一种基于梯度的优化算法 LookAhead 来更新模型参数，并使用了学习率调度策略以提高模型的训练效率和稳定性。

总体而言，本文证明了利用 Vision Transformer 和迁移学习的方法可以实现对婴儿哭声的高精度分

类识别。这一方法也可以应用于其他领域的音频分类任务。未来的工作将会继续完善婴儿哭声数据集,以提高模型的分类性能,并考虑使用其他的优化算法以进一步提高模型的训练效率。

参考文献

[1] 林浩文,张正道,张明馨,等. 一种婴儿哭声识别优化算法的研究[J]. 测控技术,2019,38(12):46-51.

[2] 国良,许鹏,沈晓燕. 基于数字信号处理器的婴儿声音识别系统的设计与实现[J]. 生物医学工程研究,2018,37(3):276-280.

[3] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2021: 12227-12236.

[4] OQUAB M, BOTTOU L, LAPTEV I, et al. Learning and transferring mid-level image representations using convolutional neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA:IEEE, 2014:1717-1724.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. Long Beach, USA:NIPS Foundation,2017: 5998-6008.

[6] LOGAN B. Mel frequency cepstral coefficients for music modeling [C]// International Symposium on Music Information Retrieval. Plymouth, USA:dblp,2000:1-6.

[7] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009:248-255.

[8] BA J, KINGMA D P. Adam: A method for stochastic optimization [C]// International Conference on Learning Representations. San Diego, USA:dblp, 2015:1-15.

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

[10] TAN Mingxing LE Q V. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR′ 2019). Long Beach, USA:IEEE,2019:6105-6114.

[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR′2016). Las Vegas, USA:IEEE,2016: 770-778.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556,2014.

[13] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995,20(3): 273-297.

[14] BOTTOU L. Large-scale machine learning with stochastic gradient descent [C]// Proceedings of COMPSTAT′2010. Princeton, USA:Springer, 2010: 177-186.

[15] REDDI S J, KALE S, KUMAR S. On the convergence of adam and beyond[J]. arXiv preprint arXiv:1904.09237,2018.