黄一舟, 崔璨, 李增. 基于 RegNet 神经网络的声纹识别[J]. 智能计算机与应用,2024,14(8):197-202. DOI:10.20169/j.issn.2095-2163.240832

基于 RegNet 神经网络的声纹识别

黄一舟[1], 崔 璨[1], 李 增[2]
(1 中国人民警察大学 研究生院, 河北 廊坊 065000; 2 中国人民警察大学 警务装备技术学院, 河北 廊坊 065000)

摘 要: 通过寻找 RegNet 神经网络的最佳宽度、神经元权重和偏置等参数的最优组合,将优化后的 RegNet 神经网络应用于语谱图的识别,实现对语谱图的特征学习和分类,进而达到对声纹进行鉴定。实验结果表明,在使用相同的 Thchs-30 数据集训练下,RegNet 网络与目前主流的卷积神经网络(CNN)和 DenseNet 识别方法相比,识别率分别高了 11.75%和 1.93%,并且达到收敛后具有更好的稳定性,相对于 DenseNet 和传统的 CNN 有更好的精确度和鲁棒性。

关键词: 声纹识别; 语谱图; RegNet; 卷积神经网络
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2024)08-0197-06

Voiceprint recognition based on RegNet neural network

HUANG Yizhou[1], CUI Can[1], LI Zeng[2]
(1 College of Graduate, China People's Police University, Langfang 065000, Hebei, China;
2 College of Police Equipment Technology, China People's Police University, Langfang 065000, Hebei, China)

Abstract: By finding the optimal combination of parameters such as the optimal width, neuron weights and bias of the RegNet neural network, the optimised RegNet neural network is applied to the recognition of speech spectrograms to achieve feature learning and classification of the speech spectrograms, and then to achieve the identification of voiceprints. The experimental results show that the RegNet network has a higher recognition rate of 11.75% and 1.93% compared with the mainstream convolutional neural network (CNN) and DenseNet, and has better stability after convergence and better accuracy and robustness compared with DenseNet and traditional CNN.
Key words: voiceprint recognition; spectrogram; RegNet; convolutional neural network

0 引 言

声纹识别的研究早期主要依赖于基于频谱分析的浅层特征,如美尔频率倒谱系数(MFCC)[1]。然而,这些浅层特征无法动态地反映语音信号的变化,存在一定的局限性。随着神经网络模型的发展,声纹特征也逐渐具有了多样性[2],从最早只具备单一的频谱特征参数逐渐具有了时间、空间化信息的结构参数[3]。在使用计算机模型识别方面,较早的声纹识别模型是高斯混合模型(GMM)[4],但这类统计模型不能充分利用声纹中的时间特征,因此应用场景较为单一,泛化性能差。进入深度学习时代,卷积神经网络[5](CNN)和 DenseNet 在声纹识别方面表现出优异的性能。不过这两种模型也存在各自的局限性,CNN 在提取空间特征方面表现出色,但其处理时间序列数据的能力有限;DenseNet 模型的计算复杂度和内存需求相较于其他模型较高,计算时往往需要消耗大量的计算资源。

本文使用调节神经网络(RegNet),通过调节网络的宽度、神经元偏置和权重等参数,以及使用交叉熵损失函数构建了用于语谱图识别的神经网络,从识别精确度和损失值两个方面衡量 RegNet 的识别效果,并比较了 RegNet 与 DenseNet 和 CNN 声纹识别精确度差异,证明了 RegNet 在语谱图声纹识别方面具有较好的识别效果,是一种有效的声纹识别方法。

1 声纹识别模型构建

1.1 RegNet 神经网络

近年来,神经网络架构搜索(NAS)[6]技术非常

基金项目: 河北省省级科技计划资助(21370302D)。
作者简介: 黄一舟(1998-),男,硕士研究生,主要研究方向:声纹识别; 崔 璨(1999-),男,硕士研究生,主要研究方向:单警定位装备在警务实战的应用。
通讯作者: 李 增(1982-),男,博士,副教授,硕士生导师,主要研究方向:警用装备信息化。Email:lizeng@cppu.edu.cn
收稿日期: 2024-02-27

受欢迎,但其对计算资源的要求也很高。基于单个网络的传统 NAS 方法存在参数调整不灵活、泛化能力差、解释性差等缺陷。NAS 通过找到神经网络某些参数组合,如深度和宽度的函数关系,来得到神经网络的最佳参数设置。典型的使用 NAS 神经网络 EfficientNet[7] 的设计思想是使用一个简单高效的复合系数,从深度、宽度和分辨率的 3 个维度来放大网络。这种方法不会像传统方法那样随机地放大网络尺寸。基于神经结构搜索技术,NAS 可以获得最佳的参数集,并计算出一定的计算成本下可以使用的最佳模型。

RegNet 也使用 NAS 技术,但与一些之前的 NAS(如:MobileNetV3、EfficientNet)不同。之前的 NAS 是使用搜索算法,在给定的设计空间中找到参数的最佳组合[8];然而,RegNet 是找到最佳的网络设计原则[9],而不仅仅是一组参数。RegNet 并非单一的网络,但也与 EfficientNet 有很多的扩展网络结构不同,其是一个由量化线性规则的设计空间[10],在相同的训练设计和每秒浮点运算次数(*FLOPs*)下,RegNet 比 EfficientNet 更准确,并且在 GPU 上比 EfficientNet 快 5 倍。RegNet 网络的结构框架如图 1 所示。

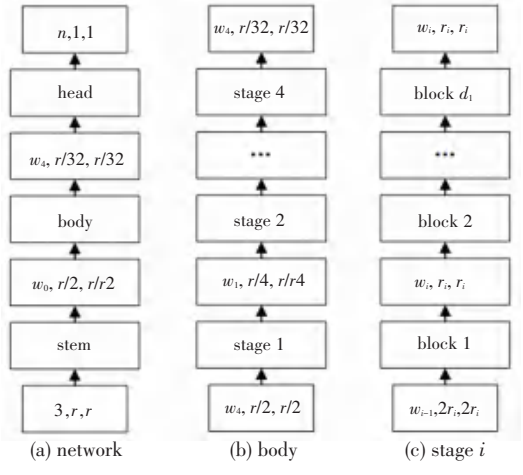


图 1 RegNet 网络结构图

Fig. 1 RegNet network structure diagram

如图 1(a)所示,RegNet 网络可以分为 stem,body 和 head 3 个部分。在 RegNet 网络结构中,body 部分为网络的主体部分,主体部分包含一系列的网络阶段(RegStage),每个 RegStage 由一系列的块堆叠而成。这些块通常采用残差结构,并带有分组卷积。每个 RegStage 中块的数量(d_i)、输出特征矩阵的通道数(w_i),以及块中每个组的宽度(g)是网络的主要参数,可以根据需要进行调整。Stem 是一个普通的卷积层(默认使用批归一化和 Relu 激活函数),其主要作用是对输入的图像进行预处理,通常包含一个步长为 2,卷积核大小为 3×3 的卷积层,用于生成初始的特征图。Head 层则负责对从主体部分提取的特征进行分类,通常包含全局平均池化(Global Average Pooling,GAP)、随机失活神经元函数(Dropout)和全连接层(Fully Connected layer,FC)。这种设计方法允许 RegNet 根据特定的任务和计算资源,灵活地调整网络的深度、宽度和分组卷积,从而实现性能的优化。如:图 1(b)中,body 部分分为 4 个阶段,类似于一个堆栈;图 1(c)中,在每一个阶段的每一步后,输入特征矩阵的高度和宽度将减少为原来的一半。一个阶段由许多块堆栈组成,每个阶段的初始块由两个快捷层和两个主要卷积层组成,其步长为 2,其余块中的卷积步长为 w_i - 1,2r_i,2r_i。

在 RegNet 中,块的主要结构是一个 1×1 卷积、一个 3×3 组卷积和一个 1×1 卷积(不包含 ReLU 函数)。在跳跃分支中,当步长为 1 时,不做任何处理,当步长为 2 时,通过 1×1 卷积执行下采样。如图 2 所示,r_i 表示分辨率,W_i 和 s 分别表示特征矩阵的宽和步长。当步长 s 等于 1 时,输入输出 r_i 不变;当 s 等于 2 时,输出 r_i 是输入的一半。W 为特征矩阵的通道(当 s = 2 时,输入为 W_i - 1,输出为 W_i;即通道会发生变化),g 表示组卷积中各组的组宽,b 代表瓶颈比,即输出特征矩阵的通道减少为输入特征矩阵通道的 1/b。

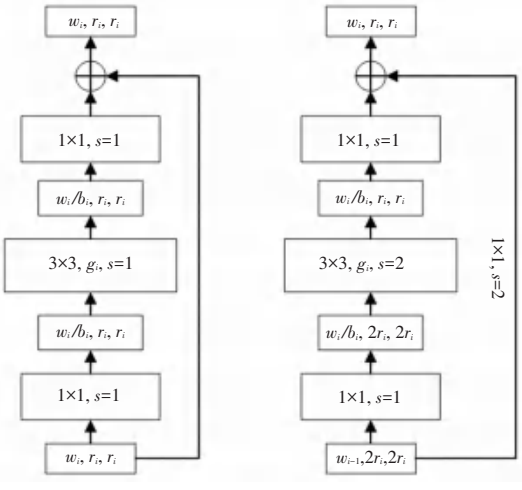


图 2 RegNet 块结构图

Fig. 2 RegNet block structure diagram

RegNet 网络的设计理念是通过调整一组关键参数来控制网络架构的分布,进而优化网络性能并加速训练过程。这些参数主要包括:每个网络阶段

(RegStage)中 Block 的数量(d_i)，每个 RegStage 中输出特征矩阵的通道数(w_i)，以及 Block 中每个 Group 的宽度。通过精细调整这些参数，RegNet 旨在优化网络性能的同时减少模型的参数数量，从而提升训练速度。本文构建的 RegNet 网络模型结构为：网络初始深度为 21 层，其中第一层是输入层，输入数据为预处理后的语谱图数据集；连接着输入层的是一系列卷积层，输入数据经过一系列卷积层，以学习和提取图像特征；卷积层之后是激活函数，使网络能够学习更复杂的特征和模式；接着是池化层，用于减小特征映射的尺寸以及减少计算量。在经过多个卷积、激活和池化层后，特征映射被展平并传递给多个全连接层；最后一个全连接层的输出被输入到一个具有与类别数量相同神经元数量的输出层中。

1.2　RegNet 损失函数的选择

在声纹识别任务中，为了使模型可以更快地达到收敛，并且通过损失函数后可以使输入的语谱图通过网络模型得到的特征图接近于原始的标签图像，保证模型有更好的准确性和鲁棒性，本文采用交叉熵损失函数(CrossEntropyLoss)作为损失函数[11]。交叉熵损失函数的表达式为

$$L = -\sum_{h,w}\sum_{c} Y_n^{(h,w,c)} \log\left(S(X_n)^{(h,w,c)}\right) \qquad (1)$$

式中：L 表示交叉熵损失函数，Y_n 表示标签图像，X_n 表示输入图像，S 表示分割图像，h、w、c 表示图像的高度、宽度和通道数。

相比于均方误差(Mean Squared Error, MSE)等其他损失函数，交叉熵损失在梯度反馈上表现更好。交叉熵损失会对那些分类错误的实例施加更大的惩罚，对于分类错误的实例，梯度将会更大，进而使得模型在训练过程中有更快的收敛速度。

2　语音信号的特征提取和标准化处理

2.1　语谱图提取

语谱图是一种二维时间-频率表示，用于分析音频信号的频谱特性[12]。语谱图可以反应声音传播过程中时间与能量强度的关系和时间特征信息，这些信息根据颜色的深浅形成了不同的纹理，在这些纹理中包含了大量的说话人的个性特征信息，根据语谱图纹理的区别，可以鉴别不同的说话人[13]。为了更好地描述音频信号的频率成分，本文采用梅尔频谱图(Mel-spectrogram)[14]作为声纹特征的提取方法。

如图 3 所示，梅尔频谱图是基于人耳听觉特性的一种语谱图表示，其主要考虑了人耳对不同频率的敏感程度。梅尔频谱图的提取过程如下：

(1)读取语音信号后，对语音信号进行预加重处理；

(2)预加重：通过对语音信号的高频部分进行加重，使得频谱趋于平坦，从而降低噪声的影响；

(3)将预加重后的信号分帧，并对每帧信号应用汉宁窗(Hanning window)以减少帧间连续性的影响

(4)对经过汉明窗处理的帧信号进行快速傅里叶变换(FFT)，得到频谱图；

(5)将频谱图转换为梅尔频谱图。将线性频率划分为梅尔尺度，然后通过应用梅尔滤波器组(Mel filter bank)将线性频谱图转换为梅尔频谱图。梅尔滤波器组采用三角形滤波器，其中心频率呈等距排列；

(6)对梅尔频谱图进行对数处理，以模拟人耳对声音响度的非线性响应。

在本文中，音频数据的采样率为 16 kHz，汉宁窗的长度为 1 024，窗口与窗口之间的重合长度为 512，执行傅里叶变换的长度与窗口长度一致为1 024，梅尔滤波器组的数量为 64，即频谱将被分为 64 个梅尔频带，分析的频率范围为 62.5 Hz 到 8 000 Hz。

图 3　梅尔频谱图

Fig. 3　Mel spectrogram

按照图 4 所示步骤，可以得到音频信号的梅尔频谱图。该方法在保留信号的时间-频率特性的同时，考虑了人耳对不同频率的敏感程度，因此在声纹识别任务中具有较好的性能。

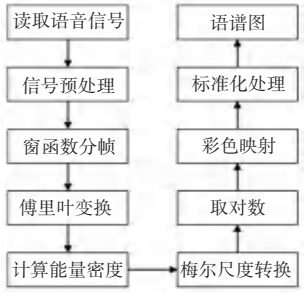


图 4　语音信号预处理流程图

Fig. 4　Voice signal preprocessing flowchart

2.2　标准化处理

由于每个音频数据的长度不一致，生成的语谱

图大小也不一致,因此首先对输入的语谱图进行预处理,将输入的语谱图图像缩放到 224×224 大小的正方形彩色图像,然后将其转化为灰度图像,最后转化为矩阵形式并标准化以匹配模型的输入需求。在实验中并未裁剪图像而是采取缩放的方法也是为了最大程度的保留语音数据的特征保证训练的效果。在将语音信号转化为梅尔频谱图之后,在输入到网络模型之前还应该将彩色的语谱图转化为灰度图像。考虑到在声纹识别任务中颜色信息对特征提取的影响较小,并且使用灰度图像可以提高模型的训练和推理的速度。由于 RegNet 的网络输入通道为三通道,因此还需要将得到的灰度值复制到其他两个通道中作为输入。此外,为了保证模型的空间不变性和防止模型的过拟合,应加入一些数据增强的方法。

在得到灰度图像之后,对于输入到神经网络的图像会进行随机旋转和裁剪大小,其中随机旋转的角度在-20°~20°之间,并且还有 50%的概率水平翻转。随机旋转是通过旋转矩阵与像素位置相乘实现的:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix} \tag{2}$$

其中, θ 表示随即旋转的角度,x 和 y 分别表示像素的水平位置和竖直位置。通过与每个像素点的相乘就可以得到旋转之后的图像。 随机裁剪则是随机选取图像的一片区域,通过调节 scale 和 ratio 两个参数来指定随机裁剪的方式。其中,scale 参数决定裁剪的面积比例,ratio 决定裁剪的宽高比,裁剪之后放大 224×224 的大小。经过以上处理流程,可以得到较为干净、稳定的梅尔语谱图特征。

3 仿真分析

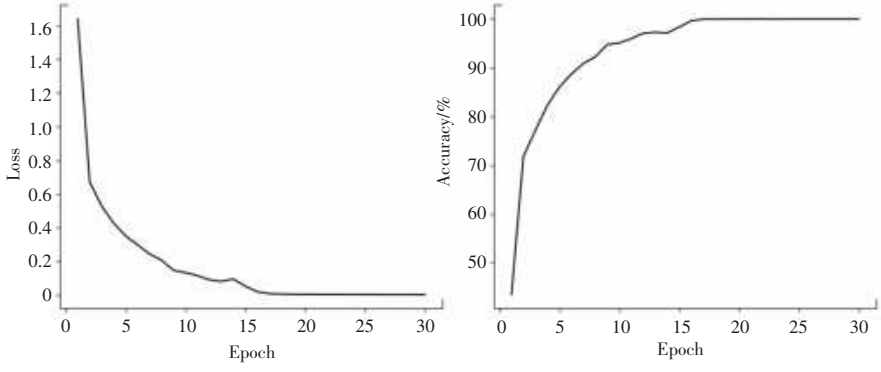
实验所用的平台为 Jupyter Notebook,开发框架为 PyTorch。实验环境为 NVIDIA GeForce RTX 3070 Ti Laptop GPU GDDR6 @ 8 GB (256 bits)。本实验使用 ReLu 激活函数,使用 Softmax 函数计算输出层的结果以便为每个类别生成概率分布。实验中设置迭代次数为 30 次。

3.1 数据集

本实验使用的标准数据集为 THCHS-30[15],该数据集是由清华大学语音和语言技术中心发布的中文数据集。音频在安静的工作空间中进行的录制,主要参与者朗读中文文本进行录音,主要的参与者大多为女性。每个音频片段大约持续 10 s,采样频率设定为 16 kHz,采样大小为 16 位。为了方便后续的实验,实验中对数据集中的数据重新进行了分类和标签,并对数据进行了筛选和统计。从中选出 44 名朗读者,共 9 869 条语音进行实验,每条语音都不相同,数据集中每个子文件夹的名称为朗读者的标签。每次实验中,都选取每个朗读者的 80%数据作为训练集,20% 数据作为测试集。即训练集共有 7 895条数据,测试集共有 1 974 条数据。

3.2 结果分析

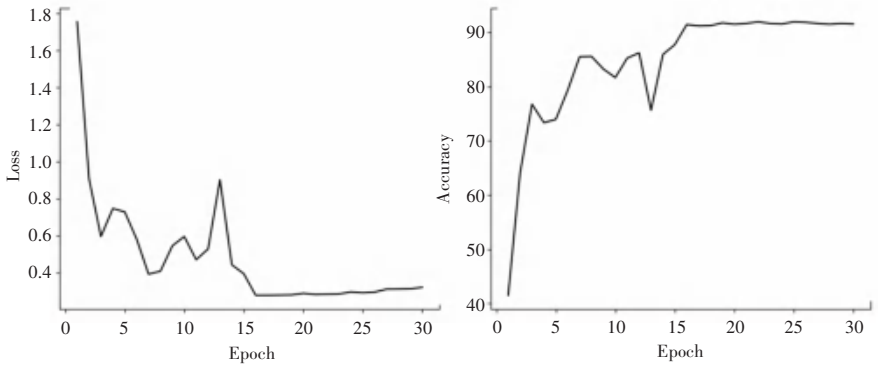
采用的性能评价指标主要包括识别正确率和损失值。识别正确率是识别正确的数据数量与总测试数量的比值,损失值是在训练和测试阶段预测的结果与实际标签之间的差异。图 5 和图 6 分别显示了 RegNet 模型在 THCHS-30 训练集和测试集上的识别准确率(Accuracy)和损失值(Loss)的变化曲线。分析发现,训练集的识别准确率在 15 轮迭代后达到收敛,稳定在 99%附近,损失值在 30 轮迭代之后达到0.000 5;测试集上的最佳识别正确率达 91.95%,损失值在 15 轮迭代之后稳定在 0.3 左右。



(a) 训练集损失率 (b) 训练集正确率

图 5 训练集上准确率和损失值的曲线图

Fig. 5 Curves of Accuracy and Loss values on the training set



（a）测试集损失率　　（b）测试集正确率

图 6　测试集上准确率和损失值的曲线图

Fig. 6　Curves of Accuracy and Loss values on the test set

从图中还可以看出，在训练过程中，RegNet 的识别准确率随着训练次数的增加而稳步上升，损失值则随着训练次数稳步下降；而在测试过程中，RegNet 的识别准确率和损失值都出现了较大的波动，主要是由于 RegNet 网络在识别任务中会自动进行宽度、神经元偏置[16]和权重的调整，导致正确率和损失值会在模型收敛之前出现动态变化。在实际识别过程中，模型的动态变化更有利于局部寻优，方便在特定条件下寻找参数的最佳组合。

为了衡量 RegNet 识别方法的优劣，将 RegNet 网络与目前在声纹识别任务中比较热门的 DenseNet 模型和传统 CNN 模型进行对比，设计了 3 组实验，在相同的数据集和数据预处理的条件下分别使用 RegNet 网络、DenseNet[17]网络和 CNN 网络进行仿真实验 。在实验 1 中，使用的是 RegNetY－064 网络，模型初始深度为 21，初始宽度为 80，每个组的大小为 24。实验 2 使用的 DenseNet 模型为 Densenet201 网络，网络深度为 201，增长率为 12，块数为 4。实验 3 使用的网络是传统的 CNN 神经网络，在这个网络中包含 3 个卷积层，一个全连接层和一个输出层，使用的激活函数为 ReLu[18]，此外还添加了一层 Dropout 层防止过拟合。本次实验中，使用的是 Adam 优化器，批量大小为 32，学习率则是在训练过程中动态调整，每 15 轮训练后学习率 $lR \times 0.1$，初始学习率 $LR = 0.001$。性能评价指标为正确率和损失值，正确率的计算方式为预测结果与真实语音的比值，损失值是通过损失函数计算的，实验中使用的是交叉熵损失函数[19]，实验结果如图 7 所示。

通过实验对比，在相同的 THCHS－30 语音数据集上，图 7 显示了 3 组实验在 30 轮迭代中达到的最高正确率比较。由图 7 可知，在 3 个实验中，DenseNet 相比 CNN 高出了 11.75% 的正确率，RegNet 相比 DenseNet 高出了 1.93% 的正确率。为了进一步比较 3 组实验的准确率，模型识别正确率与迭代次数的关系如图 8 所示。由图 8 可以看出，3 组实验都可以正常收敛，RegNet 和 DenseNet 的识别率相比于 CNN 都有明显提升，这说明 RegNet 和 DenseNet 网络在声纹识别任务中都有更好的性能。RegNet 和 DenseNet 相比，RegNet 模型达到收敛后识别率更加稳定且高于 DenseNet 模型。

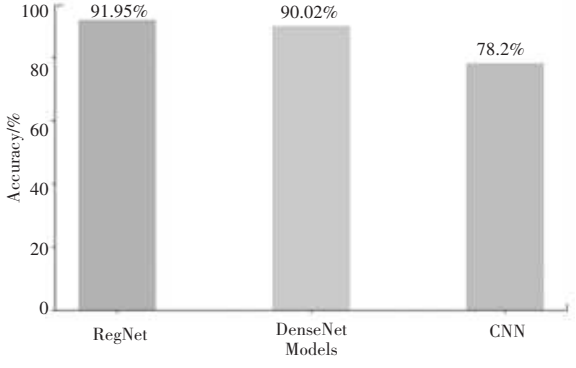


图 7　平均正确率比较图

Fig. 7　Average Accuracy comparison chart

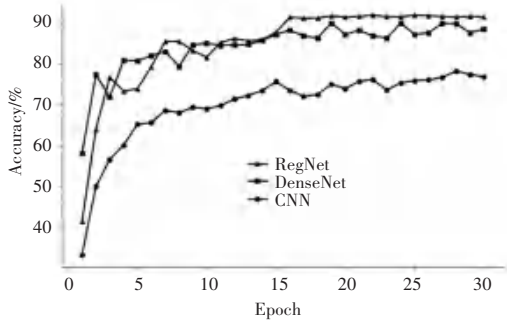


图 8　迭代次数对正确率的影响

Fig. 8　Impact of iteration number on Accuracy

4 结束语

RegNet 作为一种深度学习网络,其特性能够从语谱图中有效地提取声纹特征,从而实现精确的声纹识别。论文中提出了可以自适应改变网络结构的 RegNet 网络模型,并使用了交叉熵损失函数,并将其应用于语谱图声纹识别。最后从 RegNet 的识别精确度和对比其他方法进行研究,研究结果表明,RegNet 模型本身在声纹识别任务中有较高的正确率,在 THCHS-30 数据集中可以达到 91.95%的识别正确率。并且使用 RegNet 相比于传统的 CNN 识别方法和 DenseNet 识别方法在模型达到收敛后有更高的精确度和鲁棒性。本次实验使用开源数据集,不足以完全模拟真实环境,限制了模型的泛化能力,未来可以考虑在数据充足和硬件允许的条件下做更深层次的研究,进一步提升模型的性能。

参考文献

[1] 张玉杰,张赞. DenseNet 在声纹识别中的应用研究[J]. 计算机工程与科学,2022,44(1):132-137.

[2] TANDEL N H, PRAJAPATI H B, DABHI V K. Voice recognition and voice comparison using machine learning techniques: A survey[C]//Proceedings of 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020: 459-465.

[3] ELSKEN T, METZEN J H, HUTTER F. Neural architecture search: A survey[J]. Journal of Machine Learning Research, 2019, 20(1): 1997-2017.

[4] 刘晓璇,季怡,刘纯平. 基于 LSTM 神经网络的声纹识别[J]. 计算机科学,2021,48(S2):270-274.

[5] ZHANG C, KOISHIDA K. End-to-end text-independent speaker verification with triplet loss on short utterances[C]// Proceedings of Interspeech. IEEE,2017: 1487-1491.

[6] HU Y, JIANG X, LIU X, et al. Nas-count: Counting-by-density with neural architecture search [C]//Proceedings of Computer Vision-ECCV 2020: 16th European Conference. IEEE, 2020: 747-766.

[7] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks [C]//Proceedings of International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[8] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2020: 10428-10436.

[9] GAO Y, BAI H, JIE Z, et al. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE , 2020: 11543-11552.

[10] ZHAO B, ZHANG H, LAN X, et al. Regnet: Region-based grasp network for end-to-end grasp detection in point clouds [C]// Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 13474-13480.

[11] HO Y, WOOKEY S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling[J]. IEEE Access, 2019, 8: 4806-4813.

[12] CHAUHAN R, GHANSHALA K K, JOSHI R C. Convolutional neural network (CNN) for image detection and recognition[C]// Proceedings of 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE, 2018: 278-282.

[13] LIANG S, ZHANG R, LIANG D, et al. Multimodal 3D DenseNet for IDH genotype prediction in gliomas[J]. Genes, 2018, 9(8): 382.

[14] BIRCH B, GRIFFITHS C A, MORGAN A. Environmental effects on reliability and accuracy of MFCC based voice recognition for industrial human-robot-interaction [J]. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2021, 235(12): 1939-1948.

[15] WANG D, ZHANG X. Thchs-30: A free chinese speech corpus [J]. arXiv preprint arXiv:1512.01882, 2015.

[16] RAHAMAN N, BARATIN A, ARPIT D, et al. On the spectral bias of neural networks [C]// Proceedings of International Conference on Machine Learning. PMLR, 2019: 5301-5310.

[17] ZHU Y, NEWSAM S. Densenet for dense flow [C]// Proceedings of 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 790-794.

[18] LI Z Y, YUAN Y. Convergence analysis of two-layer neural networks with relu activation[J]. arXiv preprint arXiv: 1705.09886,2017.

[19] ZHANG Z. Improved adam optimizer for deep neural networks [C]//Proceedings of 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018: 1-2.