基于 CNN-LSTM 网络的声纹识别研究

闫 河[1,2] 董莺艳[1] 王 鹏[1] 罗 成[1] 李 焕[1]

[1](重庆理工大学计算机科学与工程学院 重庆 400054)

[2](重庆理工大学两江人工智能学院 重庆 400020)

摘 要 传统声纹识别方法过程复杂,模型识别准确率低,是声纹识别应用发展的关键问题。利用深度学习具有自主特征提取及分类的特点,结合卷积神经网络(CNN)和长短期记忆网络(LSTM),提出一种结合的网络模型学习声纹识别特征及对其进行身份认证。将原始语音转换为固定长度语谱图,顺序进入 CNN、LSTM,结合网络进行训练以及声纹特征学习。通过对比 CNN、LSTM 以及 DNN 网络,验证 CNN-LSTM 网络在声纹识别中具有较少迭代次数情况下高准确率的特性。经实验结果可以得出,语音空间特征及时序特征均是声纹识别中重要的影响因素,实验中的 CNN-LSTM 网络模型准确率达到 95.42%,损失低值达到 0.097 3。该方法有利于实际声纹识别的应用。

关键词 声纹识别 CNN-LSTM 网络 语谱图 时序特征

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.04.026

VOICEPRINT RECOGNITION BASED ON CNN-LSTM NETWORK

Yan He[1,2] Dong Yingyan[1] Wang Peng[1] Luo Cheng[1] Li Huan[1]

[1](*College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China*)

[2](*College of Artificial Intelligence, Chongqing University of Technology, Chongqing 400020, China*)

Abstract The traditional voiceprint recognition method is complex with low recognition accuracy, which is a key issue in the development of voiceprint recognition applications. In this paper, we used deep learning with autonomous feature extraction and classification, combining with convolutional neural network (CNN) and long-term and short-term memory network (LSTM). A combined network model was proposed to learn the features of voiceprint recognition and identity authentication. The original speech was converted into a fixed-length spectrogram, and sequentially entered into the combined network CNN and LSTM for training, and learning voiceprint feature. By comparing CNN, LSTM and DNN, We verified the high accuracy of the CNN-LSTM network in voiceprint recognition with fewer iterations. The experimental results show that the speech space features and time series features are important factors in voiceprint recognition. The accuracy of CNN-LSTM network model in the experiment reaches 95.42%, and the loss value is 0.0973. The method is benefical to the practical application of voiceprint recognition.

Keywords Voiceprint recognition CNN-LSTM Network Spectrogram Timing features

0 引 言

声纹识别是生物特征识别中重要的组成部分,由于声纹采集过程简单、且声音短期内具有不变的特性,能够作为身份认证的关键特征。其过程是将说话人的语音特征提取,并与原有特征进行比对,当相似度达到一定阈值后确认身份。声纹识别中的声纹辨认常应用

收稿日期:2018-11-16。国家自然科学基金项目(61173184);重庆市自然科学基金项目(cstc2018jcyjAX0694)。**闫河**,教授,主研领域:图像识别。**董莺艳**,硕士生。**王鹏**,硕士生。**罗成**,硕士生。**李焕**,硕士生。

于刑侦破案、罪犯跟踪、国防监听、个性化应用等方面,同时在证券交易、银行交易、公安取证、汽车声控锁等方面的声纹确认研究也逐渐开始。

传统声纹识别包括语音信号预处理、特征提取及模型匹配三个阶段。其中声纹特征的提取是识别过程的基础,特征表达的性能对后续识别效果影响较大,由于计算机性能的急速发展,原来受计算机内存限制的深度学习再次发展起来,学术界在声纹识别方面的讨论也渐渐从传统方法转向了深度学习方法[1-3]。在声纹特征方面,Google 提出了 d-vector[4] 特征,该特征从 DNN 网络中最后一层提取激活后的数据,进行 L2 正则化后累加,得到 d-vector 特征向量。x-vector[5] 则是从 TDNN 网络[6] 中提取的降维特征,其中 TDNN 是时延架构,其输出层能够学习长期特征,能利用 10 s 左右的语音材料,获取用户声纹信息。就目前综合特征的表达,语谱图逐渐进入了人们的视野,其能综合时间方向上的频率和语音能量的特点,形成具有综合表征意义、能代表一个人说话特征的语音频谱图。

在网络模型方面,也提出了具有针对性的声纹识别网络结构。学术界探讨较多的有 CNN 模型和 RNN 模型,其中 CNN 网络模型被用来提取语音深层次的空间特征[7];循环神经网络 RNN 用来提取语音的时序列特征,文献[8-9]对 RNN 网络在声纹识别效果优劣上进行了相关讨论。同时也有论文在海量标注数据下,使用端到端的方法,在训练速度和数据有效利用率方面进行了相关的研究[10-12]。但是由于语音信号复杂,受环境和信道的影响较大,仅用 CNN 提取声纹特征忽略了序列语音原本的序列特性,RNN 网络虽然考虑了语音的序列特征,但由于网络本身的激活函数原因,在模型训练过程中会产生梯度消失和梯度爆炸的问题,且不容易达到理想的识别效果。

目前在声纹识别方面,学术界对长短期记忆网络 LSTM(Long Short Term Memory)[14] 在声纹识别方面的相关研究较少,没有将网络长期学习的优势与声纹时序特征进行有效地结合。由于语音中通常包含具有个性声音的空间特征,和说话语段之间的时序特征,单独的网络结构无法将两种特征进行提取,本文尝试将 CNN 和 LSTM 结合,并在声纹识别的数据库中进行了有效验证,得到语谱图的空间特征和时序特征对模型效果均有影响,CNN-LSTM 网络的识别准确率和损失值均优于 CNN 和 LSTM 两个单独的网络。

1　深度学习网络架构

1.1　卷积神经网络(CNN)

CNN 是深度学习中应用广泛的网络框架之一,由 LeCun 于 2014 年提出的一种前馈神经网络。该网络在模式分类领域应用尤为突出,避免了前期对图像的复杂预处理过程,同样以语谱图作为语音数据输入网络,也可减少对语谱图的处理操作过程。

CNN 网络由两部分组成,分别是卷积层和池化层。卷积层神经元之间进行不完全连接,同样也使得网络拥有局部感受野,局部感受野在输入的图像上交叉移动,完成整幅图像的输入,构建特征的第一个隐藏层。池化是通过减少卷积层之间的连接,来降低运算复杂度。池化方法有多重多样,常用的池化方法有最大值池化方法和平均值池化方法。由于输入图像大小与卷积核相比较大,每次的卷积核移动都会产生数量较多的参数,其中包括权值和偏置。卷积神经网络中权值和偏置采用共享方法,使网络自由参数的个数减少,加快网络的计算过程,减少存储空间的占用。

CNN 网络对图像特征有着很强的学习能力,上述共享权重的方法能够减少很大的计算量。对于由语音生成的语谱图也能过图像的方式学习具有个人特性的语音特征,并进行建模。

1.2　长短期记忆网络

长短期记忆网络(LSTM)是循环神经网络(RNN)的一种升级网络,通过记忆单元解决了学习长期依赖的问题,根据之前的信息状态推断后续的信息状态,进而建立前后信息之间的联系。该网络由 Hochreiter & Schmidhuber(1997)提出,并在近期由 Alex Graves 进行了改进和推广。LSTM 的经典之处在于通过设计避免了长期依赖问题,但每个重复单元模块中的设计却精致巧妙。相比普通的 RNN,LSTM 多出三个“门”结构,分别为忘记门、输入门、输出门,对输入的信息进行不同的处理。该设计方案解决了 RNN 网络中梯度消失的问题。

LSTM 细胞中包含 1 个或多个细胞核,用来表示单元的当前状态,上述“三个门”输出分别连接乘法单元,来控制状态变化。LSTM 细胞之间的分别接受不同时刻的特征输入,通过细胞计算后,对应输出,其关键为前后连接的神经单元之间的输入关系,在接受该时刻的输入的同时,也接受上一时刻信息的输入。具体 LSTM 网络时序关系如图 1 所示。

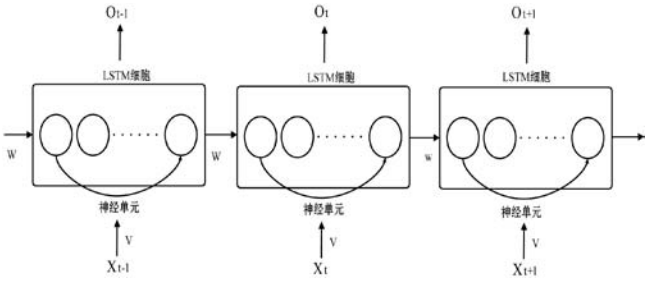

图 1 LSTM 时序关系图

2 基于 CNN-LSTM 的声纹识别网络

2.1 CNN-LSTM 网络模型

CNN 与 LSTM 均是深度学习中使用的主流算法，但对于处理不同类型的数据也各有所长。CNN 擅长提取数据局部特征，作用于空间上的抽象及泛化，能够在空间维度上提取表征能力强的高层特征，LSTM 网络能够扩展时间特征，处理具有先后顺序特征的数据信息。语音转换为语谱图的数据，以图片的形式输入网络，需要考虑其空间上的特征联系，也要考虑时间维度上关联信息。基于以上特点，本文结合 CNN 网络及 LSTM 网络的特点，采用网络串联的方式，对两个网络进行结合。得到本文使用的 CNN-LSTM 和 LSTM-CNN 网络模型，充分利用了两个网络空间、时间的表征能力，构建模型如图 2 所示。

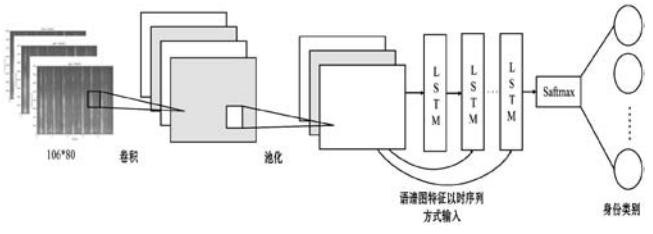

图 2 CNN-LSTM 声纹识别网络结构图

2.2 网络模型的训练

本文采取网络结构串联的方式，将 CNN 与 LSTM 网络连接，形成 CNN-LSTM 网络模型。由于语谱图可以反映说话人在各个时刻语音频谱随时间的变化，不同人的语谱图中包含个性的说话人信息，且语谱图作为图片的形式，输入深度学习网络，通过 CNN 网络能够更好地提取高表征形式的特征，故将原始语音通过分帧加窗及快速傅里叶变换后，得到可以送入网络的语谱图，其大小为 106 ×80 的 3 通道彩色图像，即输入网络的数据维度为 106 × 80 × 3。本文的实验中说话人个数为 10，说话人标签以独热编码的形式进行处理，以矩阵的形式输入网络。

本文所搭建的模型在训练前期先随机断开 20% 的神经元连接，防止由于数据维度多，网络层数少而产生过拟合现象。CNN 网络的卷积核数为 20，卷积核大小为 3 ×3，通常较小的卷积核能够对数据特征识别更加细微，且计算量较少。卷积层的激活函数选用 relu。池化层大小为 4 ×4，池化方法选择最大池化，即选择 4 ×4 范围内最大的数值作为新的池化层特征数据。经过卷积池化后的数据送入 LSTM 网络进行时序特征提取，由于 LSTM 为循环神经网络，为了防止网络内对数据的过分学习，在循环的神经单元内部及循环之间分别加入 Dropout 和 Recurrent_dropout，分别对同一 LTSM 单元中神经元之间的连接和不同循环 LSTM 单元之间的连接进行一定比例暂时断开，本文进行断开的比例为 0.2，即有 20% 的内部神经元和外部循环单元进行断开。最后接入网络的是 Softmax 全连接层，对说话人身份进行识别，实验中的人数为 10，故分类个数为 10，表 1 为上述结构参数设置的形象化表达。

表 1 CNN-LSTM 网络结构参数设置表

Relu 激活函数
Dense(10)
Recurrent_dropout(0.2)
Dropout(0.2)
LSTM(25)
最大池化(4)
Con1D(20,3)
Dropout(0.2)

3 实 验

3.1 实验平台

本实验基于 Python 的深度学习框架 TensorFlow[14] 环境下进行。实验环境如下：

（1）处理器：Inter(R)Xeon(R)CPU @2.20GHz。

（2）安装内存：32.0 GB。

（3）操作系统：Windows 7 旗舰版 64 位操作系统。

为了评估本文网络模型的有效性，使用 CNN-LSTM 网络进行声纹识别实验，实验数据采用上述同一数据集。

3.2 语音数据集和评价标准

实验采用的语音数据集为来自 Open Speech and Language Resources 的 Free ST Chinese Mandarin Corpus[15]，由 Surfingtech 提供的免费中文普通话语料库，其中包含 855 位说话人，每人包含 120 个语音片段，总计 102 600 个片段。语音采样频率为 16 000 Hz。语料库中的语音片段按照 8∶2 的比例对训练集和测试集进

行划分,其中每人的语音片段中随机抽取 80% 部分作为模型的训练集,训练个数为 96 个,剩余 20% 部分 24 个作为验证测试集,对模型的准确率等性能进行验证评价,分别对比相同训练集下不同迭代次数对准确率及损失函数的影响。

本文采用准确度 ACC(accuracy)和损失函数(Loss)作为评价标准,其中实验中使用对数损失函数。计算方法如下所示:

$$ACC = \frac{\sum_{i=1}^{n} TPi}{\sum_{i=1}^{n} (TPi + FNi)} \tag{1}$$

$$Loss = L(Y, P(Y|X)) = -\log P(Y|X) \tag{2}$$

式中:n 表示说话人数目;P_i 表示第 i 个说话人的精确度;TP_i、FN_i 分别表示第 i 个说话人中正确分类的数目和错误分类的数目。Y 表示类别正确的分类,$P(Y|X)$ 表示正确分类的概率,$Loss$ 表示为指定分类 Y 的情况下,概率越大,样本与目标值越接近,则损失越小。

3.3　实验结果及分析

本文实验中采用 10 个说话人的每人 120 个语音片段,由于选取的语音片段时常在 3 ~ 4 s 范围内,为了统一语谱图大小,将语音片段中不足 4 s 的以留白方式处理,使其统一至 4 s,然后通过傅里叶变化将语音的时序信息、频率信息和语音数据能量绘制成 106 dpi ×80 dpi 大小的语谱图,作为网络模型的输入数据。

本文通过对空间特征和时序特征提取,构建了 CNN-LSTM 模型,实验中设置的迭代次数为 20 次,为了验证本文的 CNN-LSTM 模型对声纹识别的有效性和鲁棒性,实验对比了目前在深度学习领域取得优异成果的 CNN 网络和 LSTM 网络,二者分别是模型中的单独部分与 softmax 分类器相结合组成的两个网络模型。

图 3 为实验结果图,分别为 CNN-LSTM、CNN、LSTM 在测试集中的准确率和损失函数变化。

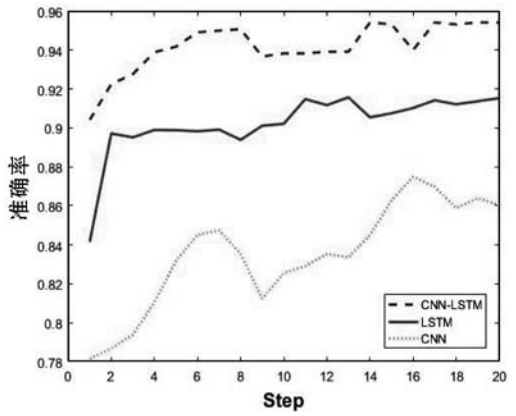

图 3　CNN-LSTM 模型及 LSTM 和 CNN 测试集准确率变化图

由图 3 可知,在测试过程中 CNN-LSTM 网络的准确率图像包含 CNN 和 LSTM 网络,即说明 CNN-LSTM 网络对于声纹识别的准确率基本高于其他两个网络。CNN-LSTM 网络的准确率随着迭代次数平稳上升,但 CNN 和 LSTM 网络在上升过程中波动较大,表明这两种模型对于数据的学习平稳性能低于 CNN-LSMT 网络。针对 CNN 和 LSTM 网络分析,LSTM 网络的准确率明显高于 CNN 网络,表明语谱图序列中时序列的特征提取性能 LSTM 网络高于 CNN 网络,且表明语谱图中具有有效的时序列特征影响声纹识别结果。实验中测试集中 CNN-LSTM 准确率 96.05%(训练集数据只表示最优结果),测试准确率 95.42%。

图 4 为 CNN-LSTM 和 CNN、LSTM 网络在测试时的损失函数变化图,由图 4 可以看出,CNN-LSTM 网络一直处于下降趋势,且变化过程平稳。最终达到训练损失函数值 0.086 9,测试损失函数 0.097 3 的低值。说明 CNN-LSTM 网络的鲁棒性较高。CNN 和 LSTM 网络的损失函数值在下降到约 0.3 左右后变化缓慢趋于平稳,且 CNN 的损失函数在下降过程中抖动频繁,表明该模型不具有较稳定的鲁棒性能,表 2 为 CNN-LSTM 及 CNN、LSTM 网络在训练和测试过程中准确率(ACC)最大值和损失函数(Loss)最小值的比较。

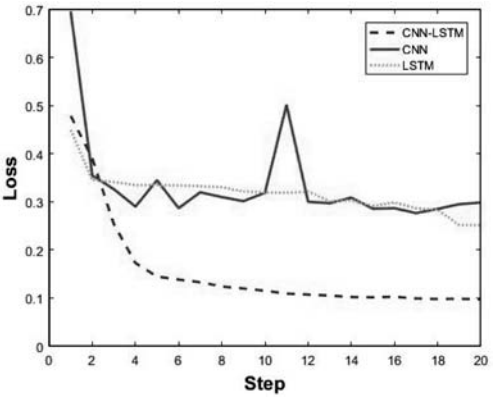

图 4　CNN-LSTM 模型及 LSTM 和 CNN 测试集损失值变化图

表 2　CNN、LSTM 和 CNN-LSTM 网络训练测试最值

网络结构	训练阶段 ACC 最大值	测试阶段 ACC 最大值	训练阶段 Loss 最小值	测试阶段 Loss 最小值
CNN	89.65%	87.48%	0.273 4	0.276 6
LSTM	92.33%	91.57%	0.241 7	0.251 4
CNN-LSTM	96.05%	95.42%	0.086 9	0.097 3

如表 2 所示,CNN-LSTM 网络在训练和测试时准确率和损失函数均优于其他两个网络结构下的数值。更加印证了语音特征中包含空间特征和时序特征,单

一特征对声纹识别影响较大,无法满足实际使用。

目前参考文献范围内,基于深度神经网络 DNN(Deep Neural Network)在声纹识别应用中取得令人瞩目的成绩,在相同数据集下,采用 DNN 网络进行对比实验,由于 DNN 网络结构深,对特征进行充分学习和网络参数进行完全训练需要大量的循环迭代次数。实验迭代结果如表 3 所示,表示迭代次数中不同网络测试集中准确率结果。

表 3 CNN、LSTM、DNN 和 CNN-LSTM 网络迭代次数对应准确率表

迭代次数	CNN/%	LSTM/%	DNN/%	CNN-LSTM/%
5	83.19	89.88	78.96	94.18
10	82.55	90.21	80.15	93.83
15	86.27	90.75	81.98	95.33
20	86.01	91.53	83.53	95.42

如表 3 所示,在迭代次数较少时,DNN 网络在以上几个网络中表现性能并不凸显,这受到网络深层结构和参数设置的限制。由于 DNN 网络在 20 次迭代中,准确率成逐渐上升趋势,但没有趋于平缓,进行实验后,发现迭代次数达到 300 左右时,准确率趋于平稳并达到 94.15%。

本文设计的 CNN-LSTM 网络能够在较少次数的迭代中达到 95.42% 的准确率,从时间效率和准确率上均优于现有 DNN 网络,故更加验证了本文提出的基于语音的空间特征和时序特征相结合的 CNN-LSTM 网络的有效性。

4 结 语

通过对比 CNN-LSTM 和 CNN、LSTM 网络模型,对声纹识别进行测试,发现 CNN-LSTM 网络模型能够很好地对语音空间特征及时间特征进行学习,对说话人身份识别认证有着较高的准确率。达到了 95.12% 的准确率和 0.097 3 的损失低值。通过与 CNN、LSTM 及 DNN 网络进行对比实验,验证了该模型在声纹识别方向的优势。本文在传统声纹识别的基础上,提出了新的高准确率识别方法,为声纹识别实际应用提供了新的思路及方法。

参 考 文 献

[1] Schmidhuber J. Deep learning in neural networks: an overview [J]. Neural Networks, 2014, 61(3): 85 - 94.

[2] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2012: 4277 - 4280.

[3] Simonyan K, Zisserman A. Very Deep Convolutional Networks for LargeScale Image Recognition [J]. Computer Science, 2014, 13(2): 120 - 131.

[4] Variani E, Lei X, Mcdermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2014.

[5] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]//Proc. InterSpeech 2017:999 - 1003.

[6] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE transactions on acoustics, speech, and signal processing, 1989, 37(3): 328 - 339

[7] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional Neural Networks for Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10):1533 - 1545.

[8] 余玲飞, 刘强. 基于深度循环网络的声纹识别方法研究及应用[J]. 计算机应用研究, 2019,36(1):153 - 158.

[9] Bhattacharya G, Alam J, Stafylakis T, et al. Deep Neural Network based Text-Dependent Speaker Recognition: Preliminary Results[C]//Odyssey 2016. 21 - 24 Jun 2016, Bilbao, Spain.

[10] Heigold G, Moreno I, Bengio S, et al. End-to-End Text-Dependent Speaker Verification[C]//Acoustics, Speech and Signal Processing(ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[11] Chowdhury F A R R, Wang Q, Moreno I L, et al. Attention-Based Models for Text-Dependent Speaker Verification [C]//ICASSP 2018—2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[12] Zhang C, Koishida K. End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances[C]// Interspeech, 2017.

[13] Greff K, Srivastava R K, Koutník, Jan, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems, 2015, 28(10):2222 - 2232.

[14] TensorFlow. 谷歌深度学习框架[EB/OL]. 2018. https://www.tensorflow.org/?hl=zh-cn.

[15] Free ST Chinese Mandarin Corpus[DB/OL]. 2016. http://www.openslr.org/38/.