计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(6):215−219 [doi: 10.15888/j.cnki.csa.007934]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 ResNet-LSTM 的声纹识别方法①

刘　勇, 梁宏涛, 刘国柱, 胡　强

(青岛科技大学 信息科学技术学院, 青岛 266061)
通讯作者: 刘　勇, E-mail: liuyong961104@sina.com

摘　要: 针对传统声纹识别方法实现过程复杂、识别率低等问题, 提出了一种基于 ResNet-LSTM 的声纹识别方法. 首先采用 ResNet 残差网络提取声纹的空间特征, 其次采用 LSTM 长短期记忆循环神经网络提取声纹的时序特征, 通过 ResNet 与 LSTM 结合的特征提取方法获得了同时包含空间特征与时序特征的深度声纹特征. 实验结果表明, 采用 ResNet-LSTM 网络的声纹识别方法的等错误率降低至 1.196%, 较基线方法 d-vector 以及 VGGNet 分别降低了 3.68% 与 1.95%, 识别准确率达到了 98.8%.

关键词: 声纹识别; ResNet-LSTM; 空间特征; 时序特征

引用格式: 刘勇,梁宏涛,刘国柱,胡强.基于 ResNet-LSTM 的声纹识别方法.计算机系统应用,2021,30(6):215–219. http://www.c-s-a.org.cn/1003-3254/7934.html

Voiceprint Recognition Method Based on ResNet-LSTM

LIU Yong, LIANG Hong-Tao, LIU Guo-Zhu, HU Qiang

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Aiming at the complex process and low recognition rate of traditional methods, this study proposes a voiceprint recognition method based on ResNet-LSTM. In this method, ResNet and LSTM are respectively used to extract the spatial and temporal features of voiceprints. Thus, the deep voiceprint features including both spatial and temporal features are obtained. The experimental results show that the equal error rate of the proposed method is 1.196%, which is 3.68% and 1.95% lower than that of the baseline methods d-vector and VGGNet, respectively, and the recognition accuracy reaches 98.8%.

Key words: voice recognition; ResNet-LSTM; spatial features; temporal features

声纹识别是生物识别技术的一种, 是计算机技术与声学、生命科学综合研究的产物之一. 与传统的身份识别技术相比, 以声纹识别、指纹识别为代表的生物识别技术具有防遗忘、防盗等特点, 并且在实际应用过程中更加方便、可靠. 生物识别技术的相关研究早已进行, 但受限于软硬件技术并不发达, 生物识别技术一直难以达到实际应用的标准. 但随着人工智能等计算机技术的高速发展, 生物识别技术取得了长足的进步, 并已广泛应用于金融、公共安全、军队国防等领域. 其中声纹识别技术由于其声纹特征采集较为方便, 在远程认证过程中具备独特优势, 并且相对于人脸识别、指纹识别等识别方法, 其对隐私的侵犯性更低更容易使用户接受, 正受到越来越多的关注.

声纹识别是指根据说话人声音中独特的声学特征自动辨别说话人身份的一种身份认证方法. 从应用场景分析可以将其分为, 说话人确认与说话人辨认两类, 其中说话人确认是一对一的判断关系, 即判断某段语音是否为指定人所发出; 而说话人辨认是一对多的选

① 基金项目: 国家自然科学基金 (61973180)
Foundation item: National Natural Science Foundation of China (61973180)
收稿时间: 2020-09-25; 修改时间: 2020-10-21; 采用时间: 2020-11-04; csa 在线出版时间: 2021-06-01

择关系, 即判断某段语音是若干说话人中哪一个所发出的. 从技术角度考虑通常可以将其分为文本相关的声纹识别方法与文本无关的声纹识别方法两类. 文本相关的声纹识别方法在训练、注册与识别阶段均需根据指定的文本内容进行发声, 该方法通常可以得到较好的识别效果但是需要用户严格按照规定文本进行发声, 灵活性较差. 文本无关的声纹识别方法没有对文本的依赖, 在应用过程中更加灵活方便, 但是其建模较为困难, 识别的准确率尚待进一步提升[1].

声纹识别一般由数据预处理、声学特征提取、模型构建、模型训练、说话人注册以及打分决策等部分组成, 流程上则可以将其分为模型训练、说话人注册以及说话人识别 3 个阶段, 如图 1 所示[2].

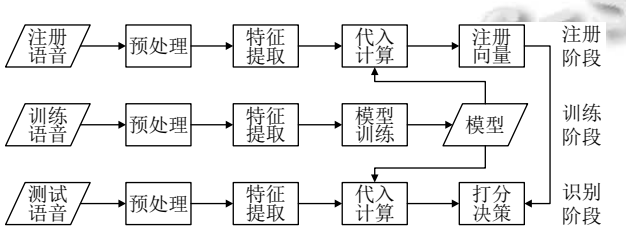

图 1 声纹识别流程图

近年来, 随着人工智能技术以及计算机软硬件理论的快速发展, 深度学习理论被广泛应用于声纹识别领域, 声纹识别的性能得到进一步提高. 如: 2014 年, Variani 等提出了利用全连接神经网络构建深度声纹特征提取网络的声纹方法[3]; 2015 年, Heigold 等提出了基于单层 LSTM 且文本相关的声纹识别方法[4]; 2017 年, Nagrani 等提出了基于 VGGNet 的声纹识别方法[5]; 2018 年, Chung 等提出了基于深度残差网络以及对比损失的声纹识别方法[6]. 通过对上述研究进行分析可以发现, 近年来提出的声纹识别方法大多存在着空间特征与时序特征只取其一, 忽略了语音片段同时包含空间特征与时序特征的问题; 例如, 文献 [4] 仅考虑了语音片段的时序特征而忽略了空间特征, 而文献 [5,6] 则仅考虑了空间特征而忽略了时序特征. 本文在文献 [4] 与文献 [6] 的基础上提出了基于 ResNet-LSTM 且与文本无关的声纹识别方法, 该方法采用 ResNet 和 LSTM 作为深度语音特征的提取网络, ResNet 部分和 LSTM 部分分别用于提取声纹中的空间特征和时序特征, 结合了 ResNet 与 LSTM 的优点, 最终的实验结果显示, 本文提出的声纹识别方法相对 d-vector 与 VGGNet 性能上均有不同程度的提升.

1 深度神经网络

1.1 卷积神经网络

卷积神经网络的概念起源于 20 世纪 60 年代, 首次提出了感受野的概念, 学者对猫的视觉皮层细胞研究发现, 每一个视觉神经元只会处理一小块区域的视觉图像, 即感受野. 20 世纪 80 年代, 日本科学家提出了神经认知机的概念, 该结构包含了 S-cells 和 C-cells 相当于卷积层和池化层, 被认为是当代卷积神经网络的原型. 1998 年, LeCun 首次提出了可以多层训练的网络结构——LeNet5, 并将 BP 算法应用至该网络结构的训练过程中, 形成了当代卷积神经网络的雏形[7]. 虽然 LeNet5 的提出是里程碑式的创新, 但是受限于计算机硬件计算能力的落后以及非常高的训练成本, 卷积神经网络一直难以媲美传统的统计学习方法, 并一直处于学术界的边缘. 直至 2012 年 Hinton 等提出了全新的 AlexNet 网络结构, 其在 AlexNet 中引入了全新的深层结构以及 Dropout 方法, 将 ImageNet 图像识别大赛的错误率降至 15%, 颠覆了图像识别领域[8]. 随后的几年中卷积神经网络在图像识别领域中得到了广泛的研究与应用, 各种优秀的卷积神经网络结构相继被提出, 如: Inception-V4[9]、VGG[10]、ResNet[11]、DenseNet[12] 等.

1.2 循环神经网络

循环神经网络是一类主要用于处理时间序列的神经网络结构, 其在语音识别、股票预测、轨迹预测等领域皆有所应用. 其主要特点在于神经元在某个时间点的输出可以再次作为神经元的输入, 这种串联结构非常适合处理时间序列问题, 可以相对保持序列数据中上下文的依赖关系. 针对循环神经网络的研究最早可追溯至上世纪 90 年代, 在长达 20 年的发展历史中诞生了多种循环神经网络结构. 如: 1997 年, Hochreiter 等提出了长短期记忆循环神经网络 (LSTM), 其在原始 RNN 的基础上做了改进, 改善了长距离的上下文依赖问题[13]; 2000 年, Gers 等提出了带有遗忘门的长短期记忆循环神经网络[14]; 2005 年, Graves 等提出了双向长短期记忆循环神经网络[15]; 2014 年, Cho 等提出了 GRU 循环神经网络等[16].

2 基于 ResNet-LSTM 的声纹识别方法

2.1 ResNet

自 2012 年 AlexNet 卷积神经网络提出以来深度

卷积神经网络已经成功应用于图像识别、语音识等多个领域, 研究人员认识到通过增加网络深度可以有效地提高卷积神经网络的性能, 但是随着网络深度的不断增加, 却出现了难以解决的梯度消失和梯度爆炸问题, 导致深度卷积神经网络在训练阶段难以得到收敛. 并且研究人员还发现随着网络深度的不断增加, 网络的退化问题愈加严重, 导致分类性能愈来愈差. 对此, He 等在 Highway 网络的基础上提出了基于残差结构的卷积神经网络——残差网络 (ResNet), 相对于 Highway 网络深度残差网络不仅缓解了深度卷积神经网络训练过程中梯度消失和梯度爆炸的问题并且大大提升了网络的性能, 在性能和训练速度上均获得了较大了提升, 成为了近年来极具影响力的一种深度卷积神经网络结构[17].

深度残差网络一般由多个残差块构成, 其中标准残差块如图 2 所示通常由卷积层 (Conv)、批量归一化层 (BN) 以及非线性激活层 (ReLU) 堆叠而成. 在普通的神经网络训练过程中, 目标是学习得到最优映射函数$H(x)$, 而在残差网络中将输入 x 直接短接至网络的输出 (跳跃连接), 此时网络将不再直接学习最优映射函数$H(x)$而是转而学习其残差$F(x)=H(x)-x$.

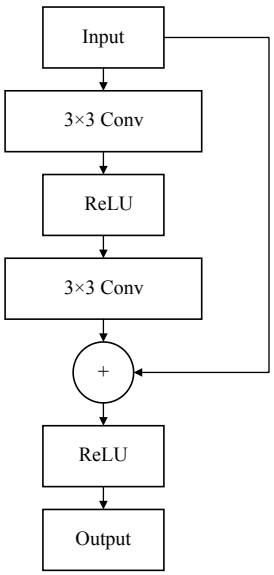

图 2 残差块结构图

2.2 LSTM

自上世纪循环神经网络提出以来在深度学习领域得到了广泛的应用, 循环神经网络的设计初衷是为了学习时间序列问题中的长期依赖性, 实践也证明循环神经网络在处理该问题上有着很好的表现, 但同时也有大量实验表明标准的循环神经网络因其迭代性将导致训练过程中出现梯度消失以及梯度爆炸问题. 为了解决此问题, Hochreiter 等提出了长短期记忆循环神经网络 (LSTM)[13], LSTM 也因此成为了实际应用中较为广泛的循环神经网络模型之一.

对比于标准循环神经网络简单的隐藏单元, LSTM 引入了门的概念并具有更复杂的隐藏单元结构, 其中隐藏单元一般由输入门 i、遗忘门 f 以及输出门 o 构成, 如图 3 所示. LSTM 对信息的存储和更新由门控部分实现, 门控可以视作一个包含了 Sigmoid 激活函数和点乘运算的全连接层. 门控操作可以公式化为:

$$g(x)=\sigma(Wx+b)$$

其中, $\sigma(x)=1/(1+\exp(-x))$ 为 Sigmoid 激活函数, 深度学习领域常见非线性激活函数之一. LSTM 中 Sigmoid 激活函数用于描述信息的通过比例, 当门的输出为 0 时, 表示没有数据通过, 当输出为 1 时表示数据全部通过[18].

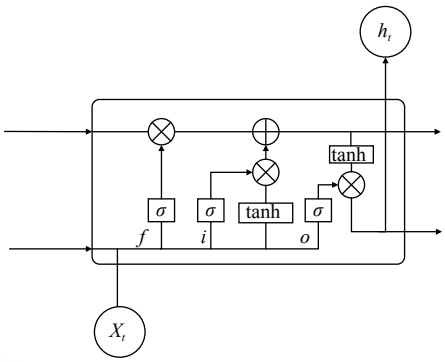

图 3 LSTM 单元结构图

2.3 ResNet-LSTM 深度特征提取网络

已知深度残差网络相对于传统的深度卷积神经网络在缓解了网络训练过程中梯度爆炸以及梯度消失问题的同时提高了网络的特征提取能力, 因此在本文中采用了深度残差网络作为声纹空间特征的提取网络. 本文采用的实验数据均为时长 1 s 的短语音片段, 虽然极短的语音片段通常难以包含具有语义上的上下文内在关系, 但时长 1 s 的语音片段经过本文的 Fbank 特征提取操作后将会获得具有 99 帧的 Fbank 特征, 其本质上依然是一个时间序列. 因此本文在提取声纹深度空间特征的同时进一步利用 LSTM 循环神经网络进行帧级别的时序特征的提取, 最终获得了同时具备空间和

时序特征的深度声纹特征.

本文的深度特征网络结构上分为两部分, 分别为 ResNet 残差网络部分以及 LSTM 长短期记忆循环神经网络部分. 其中 ResNet 残差网络部分由两个卷积层以及 6 个标准残差块构成, LSTM 长短期记忆循环神经网络部分由两个 LSTM 层构成, 详细网络结构如表 1 所示.

表 1 ResNet-LSTM 网络结构图

层	结构	步长
Conv	7×7,96	2×2
ResNet block	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times3$,96	1×1
Conv	5×5 ,256	2×2
ResNet block	$\begin{bmatrix}1\times1\\3\times3\end{bmatrix}\times3$,256	1×1
LSTM(1024)	25×1024	—
LSTM(1024)	25×1024	—
Average	1024	—
Dense	1024×1024	—
Dropout(0.2)	—	—
Dense(Softmax)	1024×340	—

2.4 等错误率

等错误率 (Equal Error Rate, EER) 是常见的模型评价指标之一, 常用于声纹识别、指纹识别、人脸识别等领域, 与错误拒绝率以及错误接受率密切相关. 对于二分类问题, 可以将其实际分类与预测分类的组合划分为表 2.

表 2 实际分类与预测分类的组合划分

实际分类	预测分类	
	正例	反例
正例	真正例(TP)	假反例(FN)
反例	假正例(FP)	真反例(TN)

错误接受率 (False Acceptance Rate, FAR) 指本不该接受的样本中接受的比例, 公式为:

$$FAR = \frac{FP}{FP+TN}$$

错误拒绝率 (False Rejection Rate, FRR) 指本不该拒绝的样本中拒绝的比例, 公式为:

$$FRR = \frac{FN}{TP+FN}$$

等错误率为判断阈值为某一值时, 错误接受率与错误拒绝率相等时的值, 此时错误接受率、错误拒绝率、等错误率三者相等.

3 实验与分析

3.1 数据选择与处理

本文采用 AISHELL-1 开源数据集作为本次实验的训练与测试数据, 该数据集包含了来自中国各地具有不同口音的 400 位说话人, 其中训练集 340 人、验证集 40 人、测试集 20 人, 总时长达到了 178 小时. 由于 ResNet-LSTM 网络中全连接层对数据输入形状有严格要求以及为了避免静音片段对网络识别能力的影响, 本文对原始数据进行了静音抑制与等长切分的预处理操作. 在未经处理的原始语音数据中存在较多的静音片段, 若不加处理对声纹识别系统将会造成严重的干扰, 影响系统的识别能力, 因此在本文中首先对原始数据进行静音抑制操作, 紧接着为了保证输入数据的大小一致对静音抑制后的语音数据进行长度 1 s 的等长切分, 在后续的模型训练以及模型测试过程中都将针对 1 s 时长的语音片段进行. 数据预处理过后紧接着是声纹特征提取操作, 本文提取了 64 维的 Fbank 特征并计算其一阶差分 (delta_Fbank) 和二阶差分 (delta_delta_Fbank), 按照 Fbank、delta_Fbank、delta_delta_Fbank 的顺序对其进行堆叠, 形成一个类似于彩色图片的三通道矩阵, 最终获得的输入数据形状为$64\times99\times3$.

3.2 实验设置

本文采用了具有 NVIDIA GTX1080Ti 高性能显卡的专业服务器, 并搭建了包括 TensorFlow-GPU 1.15.0、Keras 2.3.1、CUDA 10.0.130、cuDNN 7.6.5 的开发环境. 实验中, 在训练阶段采用交叉熵损失作为代价函数, 以及动量为 0.99、初始学习率为 0.005、衰减率为 0.0001 的随机梯度下降, 训练的总迭代次数 (epochs) 为 40, batch_size 为 32. 在说话人注册阶段, 每个人随机选取了 5 个语音片段并取其均值作为说话人注册向量. 在测试阶段, 计算待识别语音的深度说话人嵌入与注册向量之间的余弦相似度作为相似性评分, 评分越高则判断两段语音的声纹越相似.

3.3 实验结果与分析

本文以 D-vector、VGG 为本文的基线方法, 并针对本文提出的网络结构进行了消融实验, 以验证残差网络与 LSTM 循环神经网络结合的网络结构的积极作用. 本文在实验中采用等错误率 (EER) 作为本次实验的评价指标, 并利用 DET 曲线可视化比较各模型性能的差异, 详细对比了模型间的等错误率 (表 3 所示) 以及 DET 曲线 (图 4 所示).

表 3 测试集等错误率

方法	EER
d-vector[3]	0.048 71
VGG[5]	0.031 44
ResNet	0.018 66
LSTM	0.028 54
ResNet-LSTM	0.011 96

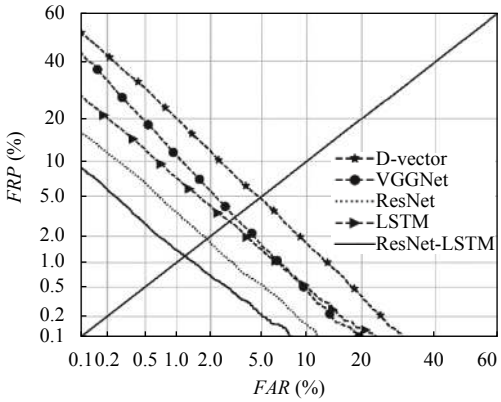

图 4 DET 曲线

由表 3 以及图 4 可以得到在声纹识别问题中, 本文提出的基于 ResNet-LSTM 的声纹识别方法明显优于基线方法中的声纹识别方法, 等错误率降低至 1.196%, 相对于对比实验中的各种声纹识别方法下降了 0.67~3.6%.

4 结语

本文提出了一种基于 ResNet-LSTM 的声纹识别方法, 该方法首先采用深度残差网络进行声纹空间特征的提取操作, 其次利用 LSTM 循环神经网络进行时序特征的提取, 结合了卷积神经网络与循环神经网络的优点. 本文通过实验证明了该方法的有效性, 与基线方法中的声纹识别方法相比, 本文提出的声纹识别方法大大降低了声纹识别的等错误率, 提高了声纹识别的准确率. 后续将进一步研究特征融合、模型融合等方法, 进一步提高声纹识别方法的识别性能.

参考文献

1 郑方, 李蓝天, 张慧, 等. 声纹识别技术及其应用现状. 信息安全研究, 2016, 2(1): 44–57.

2 吴明辉, 胡群威, 李辉. 一种基于深度神经网络的话者确认方法. 计算机应用与软件, 2016, 33(6): 159–162. [doi: 10.3969/j.issn.1000-386X.2016.06.039]

3 Variani E, Lei X, McDermott E, *et al*. Deep neural networks for small footprint text-dependent speaker verification. Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy. 2014. 4052–4056.

4 Heigold G, Moreno I, Bengio S, *et al*. End-to-end text-dependent speaker verification. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China. 2016. 5115–5119.

5 Nagrani A, Chung JS, Zisserman A. VoxCeleb: A large-scale speaker identification dataset. Proceedings of Interspeech 2017. Stockholm, Sweden. 2017. 2616–2620.

6 Chung JS, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. arXiv preprint arXiv: 1806.05622, 2018.

7 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

8 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2012. 1097–1105.

9 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4278–4284.

10 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2015.

11 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2015. 770–778.

12 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

13 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

14 Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000, 12(10): 2451–2471.

15 Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5–6): 602–610.

16 Chung J, Gulcehre C, Cho KH, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv: 1412.3555, 2014.

17 郭玥秀, 杨伟, 刘琦, 等. 残差网络研究综述. 计算机应用研究, 2020, 37(5): 1292–1297.

18 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述. 计算机应用, 2018, 38(S2): 1–6, 26.