基于多尺度残差深度卷积神经网络的语音识别

刘 虹 袁三男
(上海电力大学电子与信息工程学院 上海 200090)

摘 要 针对卷积神经网络在连续语音识别中识别性能较差的问题,提出多尺度残差深度卷积神经网络的语音识别的算法,并结合联结时序分类算法,构建端到端中文语音识别系统。将多尺度学习和残差机制以及空洞卷积引入到神经网络中,摆脱序列建模对长短时记忆神经网络的依赖,提高模型的训练速度,增强语音识别的抗噪声干扰性。实验表明,与双向长短时记忆模型(BLSTM)、深度卷积神经网络模型(DCNN)和卷积神经网络-长短时记忆模型(CNN-LSTM)相比,该模型的字错误率 WER(Word Error Rate)分别降低了 9%、5% 和 3% 左右,且在噪声环境下的识别率也优于传统的语音识别系统。

关键词 语音识别 多尺度 卷积神经网络 端到端

中图分类号 TP3 TN912.34 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.044

SPEECH RECOGNITION BASED ON MUTI-SCALE RESIDUAL DEEP CONVOLUTIONAL NEURAL NETWORK

Liu Hong Yuan Sannan
(*School of Electronics and Information Engineering*, *Shanghai University of Electric Power*, *Shanghai 200090*, *China*)

Abstract To solve the problem of poor performance of convolutional neural networks in continuous speech recognition, this paper proposes an algorithm based on a multi-scale residual deep convolutional neural network, and constructs an end-to-end speech recognition system for Chinese, by integrating connectionist temporal classification into the algorithm. The multi-scale learning, residual mechanism, and dilated convolution were introduced into the neural network to eliminate the dependence of sequence modeling on LSTM, improve the training speed of the model, and enhance the anti-noise interference of speech recognition. Experiments show that compared with BLSTM, DCNN and CNN-LSTM, the WER of this model is reduced by 9%, 5% and 3% respectively, and the recognition rate in noisy environment is better than that in traditional speech recognition system.

Keywords Speech recognition Multi-scale Convolutional neural network End-to-end

0 引 言

自动语音识别(Automatic Speech Recognition, ASR)技术是一种将人类语音转换成文字的技术,语音识别的任务主要有声纹识别、关键词识别、语言辨识、连续语音识别[1]。语音识别属于模式识别,主要通过模型算法识别声音向量,即语音的特征参数,最终给出识别结果,而语音识别的最终目标是能够在不同环境下,依旧能准确地识别出说话人说的内容[2]。早期语音识别的声学模型主要是混合高斯隐马尔可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM),但是该模型在嘈杂环境下的识别效果较差,无法满足现代自动语音识别的要求[3]。从 2009 年开始,深度学习的发展使得大词汇量的语音识别成为可能。基于深度神经网络(Deep Neural Network, DNN)的系统取代了 GMM-HMM 模型成为了主流语音识别的模型[4],但由于模型本身的缺陷,语音识别的识别率难以继续得到提高。由于语音是上下文相关的,因此长短时记忆神经网络(LSTM)的出现大大提高了语音识别的准确度,LSTM 有着强大的序列建模能力[5]。同时,Graves 等[6]将联结时序分类技术引入到循环神经网络

收稿日期: 2019-07-01。**刘虹**,硕士生,主研领域:深度学习,语音识别。**袁三男**,副教授。

的训练中,完成了序列的输入和输出自动对齐的任务。目前识别率较好的语音识别的模型主要将 CNN 和 LSTM 结合到一起,CNN 用来提取特征,LSTM 对序列建模,但是 CNN 的层数一般是两到三层,表达能力有限,提取上下文的相关性主要还是依靠 LSTM 网络。但是 LSTM 的记忆模块较小,神经网络的训练速度很慢,且实时性较差。

本文提出一种基于多尺度残差深度卷积神经网络的语音识别模型,由于卷积神经早期用在语音识别中主要是提取特征,卷积层数也较少,限制了卷积神经网络对语音识别的表达能力。因此,本文结合多尺度分析和深度残差网络,提取语音特征图中不同尺度的特征,将这些特征进行融合,最后结合联结时序分类算法构建端到端的中文语音识别模型,摆脱序列建模对 LSTM 的依赖。

1 多尺度残差深度卷积网络

随着激活函数、批量归一化和残差机制等各类算法的出现,卷积神经网络向着更深层次的方向发展,图像识别的准确率得到了进一步的提升[7]。但在语音识别中,目前主流神经网络框架中的卷积层的层数只有几层,序列建模仍然依赖 LSTM 网络,限制了卷积神经网络在语音识别领域中的发展。

1.1 卷积神经网络

CNN 主要由卷积层、池化层、全连接层、激活函数组成,卷积层用来提取特征参数,池化层对特征图进行压缩,而全连接层充当分类器的作用[8]。池化层又分为最大池化和平均池化,最大池化采用特征图的局部最大值达到特征降维的目的,在有噪声的语音中,相邻帧的时频图的局部最大值通常为语音,局部最小值为噪声,最大池化层会对含噪语音的时频图进行筛选,提高语音的分辨率,降低噪声,因此本文采用最大池化层[9]。CNN 的激活函数一般使用线性整流函数(Rectifiler Linear Unit,ReLU)。批量归一化(Batch Normalization,BN)技术的出现有效地解决了神经网络内部协变量转移的问题,大大加快了神经网络的训练速度[10]。随机失活层(Dropout)使网络的泛化性能得到提高,随机响应网络的节点,保证了网络的稀疏性[11],本文同时引入空洞卷积,在相同的卷积核大小的情况下,空洞卷积通过改变空洞率的大小来改变感受野的大小,网络的参数量不变的同时,又获得更多的上下文信息,空洞卷积实际的卷积核大小计算公式如下:

$$K = k + (k-1)(r-1) \tag{1}$$

式中:k 为原始卷积核的大小;r 为扩张率;K 为空洞卷积的实际感受野大小。

1.2 残差学习机制

批量归一化和 Dropout 层的出现加深了 CNN 的深度,通常网络越深,训练精度越高。但是随着网络层数的增加,网络参数变得难以优化[12],训练精度反而会下降。深度残差网络是 2015 年提出的深度卷积网络,其特点是简单高效,并能有效地解决网络深度变深以后的网络性能退化的问题[13]。残差网络通过学习输入到输出的目标函数与原输入的残差量,将残差量与原始输入量相加,得到最终的目标映射函数,若输入变量为 x,目标输出的实际映射为 $H(x_l)$,则残差映射 $F(x_l, W_l)$ 可以定义为:

$$F(x_l, \boldsymbol{W}_l) = H(x_l) - x_l \tag{2}$$

式中:x_l 为 l 层的输入量;$\boldsymbol{W}_l$ 为 l 层的权重矩阵。通过“捷径连接”的方式,直接把输入 x_l 传到输出作为初始结果,输出结果为 $H(x_l) = F(x_l, \boldsymbol{W}_l) + x_l$,当 $F(x_l, \boldsymbol{W}_l) = 0$ 时,$H(x_l) = x_l$。

1.3 多尺度特征

语音当前的状态,与前后的状态都有关,网络层数越多,丢失的细节信息越多,因此本文引入多尺度特征。图 1 为一段纯净语音的时频图,图 2 为加了噪声的语音时频图,两幅图所表示的语音内容相同,横向为时间轴,纵向为频率轴,该段语音有 16 s,时频图的时间轴较长。时频图反映了语音的信号强度在不同频段内随时间的变化情况。不同频率中颜色深的地方随着时间的推移,延长成声纹,由图 1 可以看出,语音信号的能量大多集中在低频,高频能量较少,但高频能量中包含很多语音的细节部分,这些细节部分也会影响语音识别的结果。由图 2 可知,在噪声背景下,语音时频图的纹理受到了干扰,但是高低频段某些纹理特征和轮廓信息依旧存在,因此模型既要能提取到细节信息,又要提取整体的轮廓信息。本文采用不同大小的卷积核以及不同空洞率的空洞卷积获取语音信号的细节信息和上下文相关性,卷积核越大,感受野越大,并对语音的时间维度和频率维度建模。图 3 中的 scale1 和 scale2 为两个多尺度子空间,scale1 的卷积核大小为 3,scale2 的卷积核大小为 5,scale1 和 scale2 各包含两个残差网络,Conv 表示卷积层,Max_pool 表示最大池化层,箭头所示即为残差结构,残差网络采用“捷径连接”的方式,相同的特征图在两个不同尺度空间下会有不同的表达形式,达到信息互补的目的。因此将这两个尺度空间融合,得到具有较好语义能力特征参数,从低层往高层逐层提取特征,得到全局信息,既可以得到相邻帧之间的相关性,也可以获取不相邻帧之间的相关性。若 scale1 网络的输出为 $f^{s1}(x)$,scale2 网络的

输出为$f^{s2}(x)$,则融合后网络输出为$f^{s1}(x)+f^{s2}(x)$。

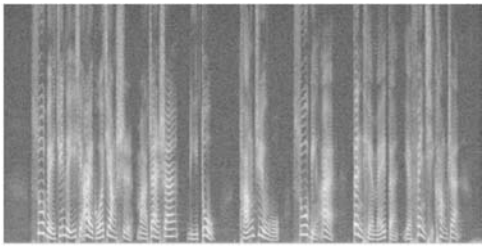

图 1　语音信号的时频图

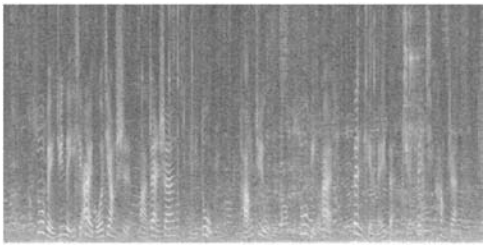

图 2　含噪语音的时频图

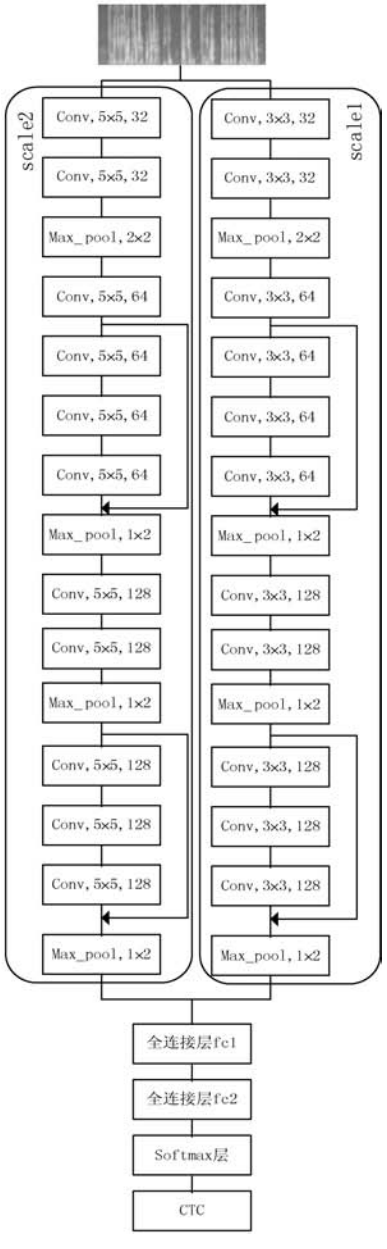

图 3　多尺度残差深度神经网络的语音识别模型

2　连接时序 CTC 的应用

CTC 通常和 LSTM 以及 RNN 一起建模,目前主流的语音识别模型都含有 LSTM 或者 RNN,但 LSTM 和 RNN 的训练受到计算机计算能力和内存的限制,训练周期较长,不利于参数调整和优化,CNN 可借助其强大的并行运算能力大大缩短训练周期。语音识别中常用 HMM 进行帧对齐,将每一帧数据对齐到 HMM 状态下[14],这种强制对齐的方法并不合理,因为语音中静音状态并没有对应的标签,而 CTC 不需要预先将输入和输出强制对齐。

在给定输入序列下 X,经过 Softmax 层输出之后,网络的输出为 $P(q_t \mid X)$,q_t为 t 时刻的输出,则标签序列 l 为所有路径概率总和为:

$$P(l \mid X) = \sum_{l \in \Gamma(q_t)} \prod_{t=1}^{T} p(q_t \mid X) \tag{3}$$

式中:$\Gamma(q_t)$为标签序列的多对一的映射。因为同一个标签序列可能有多条路径与其对应,所以需要去掉路径中的重复的标签和空白标签。$q_t \in A, t = 1,2,\cdots, T$,$A$ 为加了空白标签“—”的标签集合,输出序列中概率最大的标注序列为:

$$l^* = \arg\max_{l} P(l \mid X) \tag{4}$$

CTC 的损失函数是所有标签的负对数概率之和,可以通过反向传播训练 CTC 网络:

$$H(x) = -\log P(l \mid X) \tag{5}$$

本文所用到的模型框架如图 3 所示,输入为语音的时频特征(500 × 250),分别输入 scale1 和 scale2 两个尺度空间,提取不同尺度的特征,两个尺度空间融合后,将融合后的特征参数输入全连接层。全连接层共 2 层,第一层有 256 个网络节点,第二层网络节点的个数为 2 667 个,加上了一个空白字符。通过 Softmax 层连接 CTC 模型,学习率为 0.001,batchsize 为 8,所有卷积层和池化层后都添加了 dropout 层。CTC 在海量训练数据的条件下不再需要语言模型,因此本文的模型直接以汉字为建模单元,构建端到端的语音识别模型。

3　实　验

3.1　实验环境和数据

本文基于 THCHS30 进行实验,THCHS30 数据库

是由清华大学发布的中文语音库,有 35 小时的语音数据,数据库内容大部分来自新闻语料,全部为中文普通话,每个音频 16 s 左右。将该数据集中 10 000 句作为训练集,893 句作为验证集,2 495 句作为测试集。该数据集还提供了含有噪声的语音数据,噪声环境主要是咖啡馆、汽车内噪声和白噪声,可用于模型在不同噪声环境下的性能测试。实验采用的操作系统是 Ubuntu 16.04,GPU 为 Rtx2080ti,实验平台为 Tensorflow 3,Python 3.6。

3.2 实验结果及分析

(1) 不同模型的对比。将本文模型与 BLSTM-CTC 模型、DCNN 模型、CNN-LSTM 模型进行对比。DCNN 模型即 scale1 和 scale2。

BLSTM-CTC 模型的输入特征为 600 × 39 维的二阶差分 MFCC 特征参数,其中 BLSTM 声学模型包含 3 层隐藏层,第一层和第二层包含共 256 个前向和后向的 LSTM 单元,第三层为 512 个前向和后向的 LSTM 单元。解码单元 CTC 模型,batch size 设置为 8,学习率为 0.001。

DCNN1 模型的结构采用 scale1 尺度空间的结构,卷积核大小为 3,经过两个残差网络和最后一层池化层后,输入三层全连接层,第一层全连接层有 256 个网络节点,第二层有 512 个神经网络单元,第三层有 2 667 个网络节点,batch size 为 8,学习率为 0.001。DCNN2 模型采用 scale2 尺度空间的结构,卷积核大小为 5,全连层结构与 DCNN1 结构一致。

CNN-LSTM 模型结构由三层卷积、三层池化层、两层 BLSTM 隐藏层、一层全连接层组成。输入特征为 600 × 39 的二阶差分 MFCC 特征参数。卷积核大小为 3,池化层选择 1 × 3,只对频率维度进行池化。第一层隐藏层有 256 个网络节点,第二层有 512 个神经网络单元,全连接层有 2 667 个节点。batch size 为 8,学习率为 0.001。

本文提出的语音识别模型的识别率比单一尺度空间、BLSTM 网络和 CNN-LSTM 网络的都高,相对于 BLSTM 模型,验证集和测试集的识别率都提高 9% 左右,相对于 DCNN1 和 DCNN2 模型,识别率分别提高 5% 和 10% 左右,相对于 CNN-LSTM 网络,识别率提高 3% 左右。不同模型下的语音识别率如表 1 所示,只含 CNN 的语音识别模型每轮的训练时间比 BLSTM 模型缩短 6 倍左右,比 CNN-LSTM 模型缩短 4 倍左右。

表 1 不同模型下的语音识别率

网络模型	训练周期 /(min · epoch^{-1})	验证集 /%	测试集 /%
BLSTM	35	15.25	19.48
DCNN1	9	12.66	16.84
DCNN2	7	16.39	20.73
CNN-LSTM	27	8.63	14.72
Multi-scale DCNN	6	6.77	11.89

(2) 不同模型的抗噪声性能对比。本文同时还验证不同模型的抗噪声性能,数据集中包含咖啡馆噪声(cafe),汽车噪声(car)和白噪声(white),信噪比为 0 dB。将这些噪声加入待识别的语音中,实验结果如表 2 所示,CNN 有一定的抗噪声性能,而本文提出的 multi-scale DCNN 模型的抗噪声性能比 BLSTM 网络、DCNN 网络和 CNN-LSTM 模型都要好,更具有实用性。

表 2 不同噪声类型下的语音识别率 %

噪声类型	模型	验证集	测试集
cafe	BLSTM	19.20	22.88
	DCNN1	15.54	19.66
	CNN-LSTM	11.86	16.22
	Multi-scale DCNN	8.39	14.73
car	BLSTM	23.99	27.83
	DCNN1	18.24	21.66
	CNN-LSTM	17.33	19.82
	Multi-scale DCNN	15.27	17.33
white	BLSTM	30.57	32.64
	DCNN1	23.69	27.34
	CNN-LSTM	21.99	24.51
	Multi-scale DCNN	19.02	22.25

(3) 低信噪比下的识别率变化。本文还对比不同噪声在低信噪比下的 BLSTM、DCNN1、CNN-LSTM 模型和本文模型的误码率。由图 4 - 图 6 可知,在低信噪比下,本文提出的多尺度残差深度神经网络比 BLSTM 网络的抗噪声性能更加稳定,噪声越强,BLSTM 网络的识别率较差并且识别率下降更快,不利于实际生活中的应用。而 DCNN 网络和 CNN-LSTM 网络的抗噪声性能比 BLSTM 网络好,在 cafe 和 car 噪声下,变化相对平缓,但是在白噪声下,识别率也下降较快。因此,本文模型具有更好的鲁棒性。

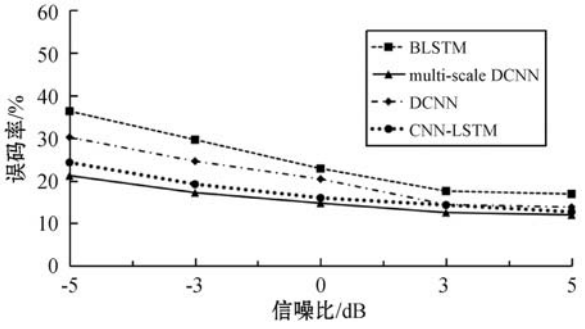

图 4　不同信噪比下不同模型的误码率(cafe)

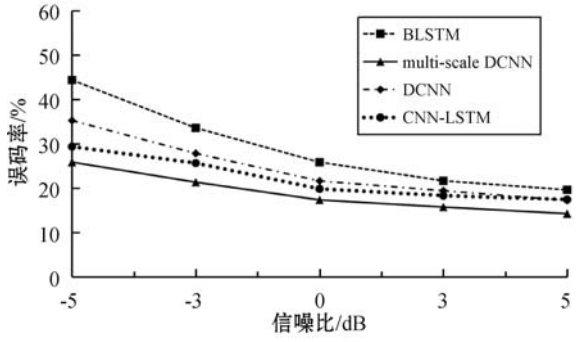

图 5　不同信噪比下不同模型的误码率(car)

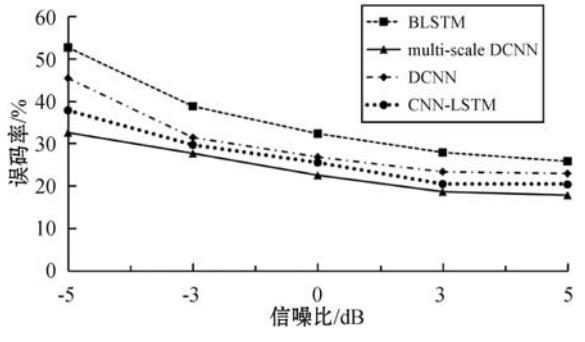

图 6　不同信噪比下不同模型的误码率(white)

4　结　语

本文针对传统语音识别模型训练时间较长以及抗噪声性能较差的问题,提出多尺度残差深度神经网络的端到端语音识别模型。该模型不仅加快网络的训练速度,而且提高语音识别的识别率和抗噪声性能。为了提取更多的上下文信息,还引入空洞卷积和多尺度特征,增大感受野,提升网络的泛化性能,降低过拟合的概率。实验证明,该模型相对于传统的语音识别的模型,有更好的稳定性和实用性。

参　考　文　献

[1] 俞栋,邓力. 解析深度学习:语音识别实践[M]. 电子工业出版社,2016.

[2] Arel I, Rose D C, Karnowski T P. Deep machine learning-a new frontier in artificial intelligence research[J]. IEEE Computational Intelligence Magazine,2010,5(4): 13 – 18.

[3] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012,29(6):82 – 97.

[4] 金超,龚铖,李辉. 语音识别中神经网络声学模型的说话人自适应研究[J]. 计算机应用与软件,2018,35(2):200 – 205,266.

[5] Hayashi T, Watanabe S, Toda T, et al. Duration-controlled LSTM for polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2017, 25(11):2059 – 2070.

[6] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]//31st International Conference on Machine Learning, 2014.

[7] Li Y, Hu J, Zhao X, et al. Hyperspectral image super-resolution using deep convolutional neural network[J]. Neurocomputing,2017,266:29 – 41.

[8] 田熙燕,徐君鹏,杜留锋. 基于语谱图和卷积神经网络的语音情感识别[J]. 河南科技学院学报(自然科学版),2017,45(2):62 – 68.

[9] 袁文浩,孙文珠,夏斌,等. 利用深度卷积神经网络提高未知噪声下的语音增强性能[J]. 自动化学报,2018,44(4):751 – 759.

[10] Laurent C, Pereyra G, Brakel P, et al. Batch normalized recurrent neural networks[C]//2016 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[11] Poernomo A, Kang D K. Biased dropout and crossmap dropout: Learning towards effective dropout regularization in convolutional neural network[J]. Neural Networks, 2018, 104: 60 – 67.

[12] 高净值,刘祎,张权. 改进深度残差卷积神经网络的 LDCT 图像估计[J]. 计算机工程与应用,2018,54(16):203 – 210.

[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition,2016.

[14] 周雁,赵栋材. 基于 HMM 模型的藏语语音合成研究[J]. 计算机应用与软件,2015,32(5):171 – 174.

(上接第 223 页)

[15] Song W S,Li C G,Wang C M,et al. Force/moment isotropy of 8/4 – 4 parallel six-axis force sensor based on performance atlases[J]. Transactions of Nanjing University of Aeronautics and Astronautics,2018,35(6):110 – 118.

[16] Xu F Y,Yang Z,Jiang G P. An analyzing and experimental method based on the resultant motion signals for SCARA manipulator joints[J]. High Technology Letters,2017,23(3): 279 – 285.