工程科学学报，第 37 卷，第 9 期：1212－1217，2015 年 9 月
Chinese Journal of Engineering，Vol. 37，No. 9：1212－1217，September 2015
DOI：10.13374/j.issn2095－9389.2015.09.015；http://journals.ustb.edu.cn

基于卷积神经网络的连续语音识别

张晴晴✉，刘　勇，潘接林，颜永红

中国科学院语言声学与内容理解重点实验室，北京 100190
✉ 通信作者，E-mail：zhangqingqing@hccl.ioa.ac.cn

摘　要　在语音识别中，卷积神经网络（convolutional neural networks，CNNs）相比于目前广泛使用的深层神经网络（deep neural network，DNNs），能在保证性能的同时，大大压缩模型的尺寸．本文深入分析了卷积神经网络中卷积层和聚合层的不同结构对识别性能的影响情况，并与目前广泛使用的深层神经网络模型进行了对比．在标准语音识别库 TIMIT 以及大词表非特定人电话自然口语对话数据库上的实验结果证明，相比传统深层神经网络模型，卷积神经网络明显降低模型规模的同时，识别性能更好，且泛化能力更强．

关键词　卷积神经网络；连续语音识别；权值共享；聚合；泛化性
分类号　TN912.34

Continuous speech recognition by convolutional neural networks

ZHANG Qing-qing✉，*LIU Yong*，*PAN Jie-lin*，*YAN Yong-hong*

Key Laboratory of Speech Acoustics and Content Understanding，Chinese Academy of Sciences，Beijing 100190，China
✉ Corresponding author，E-mail：zhangqingqing@hccl.ioa.ac.cn

ABSTRACT　Convolutional neural networks（CNNs），which show success in achieving translation invariance for many image processing tasks，were investigated for continuous speech recognition．Compared to deep neural networks（DNNs），which are proven to be successful in many speech recognition tasks nowadays，CNNs can reduce the neural network model sizes significantly，and at the same time achieve even a better recognition accuracy．Experiments on standard speech corpus TIMIT and conversational speech corpus show that CNNs outperform DNNs in terms of the accuracy and the generalization ability．

KEY WORDS　convolutional neural networks；continuous speech recognition；weight sharing；pooling；generalization

语音识别是人机交互的一项关键技术，在过去的几十年里取得了飞速的进展．传统的声学建模方式基于隐马尔科夫框架，采用混合高斯模型（Gaussian mixture model，GMM）来描述语音声学特征的概率分布．由于隐马尔科夫模型属于典型的浅层学习结构，仅含单个将原始输入信号转换到特定问题空间特征的简单结构，在海量数据下其性能受到限制．人工神经网络（artificial neural network，ANN）是人们为模拟人类大脑存储及处理信息的一种计算模型．近年来，微软利用上下文相关的深层神经网（context dependent deep neural network，CD－DNN）进行声学模型建模，并在大词汇连续语音识别上取得相对于经鉴别性训练 HMM 系统有句错误率相对下降 23.2% 的性能改善[1]，掀起了深层神经网络在语音识别领域复兴的热潮．目前包括微软、IBM 和 Google 在内的许多国际知名语音研究机构都投入了大量的精力开展深层神经网络的研究[1－3]．

实际上，人工神经网络的应用非常广泛，种类也多种多样．在文本、图像识别中，另一种更为有效的人工神经网络结构被普遍使用：卷积神经网络（convolution－

收稿日期：2014－05－08
基金项目：国家自然科学基金资助项目（11161140319，91120001，61271426）；中国科学院战略性先导科技专项（XDA06030100，XDA06030500）；国家高技术研究发展计划资助项目（2012AA012503）；中国科学院重点部署项目（KGZD－EW－103－2）

al neural networks, CNNs)[4]. 卷积神经网络来源于20世纪60年代对于猫脑皮层神经元的研究,它是一种多阶段全局可训练的人工神经网络模型,可以从经过少量预处理,甚至原始数据中学习到抽象的、本质的和高阶的特征,在车牌检测、人脸检测、手体字识别、目标跟踪等领域得到了广泛的应用,是机器学习、计算机视觉等领域研究的热点. 最近的研究表明,卷积神经网络在一些计算机视觉任务上取得了很好的结果,比如在手体字数据集和德国交通信号数据集上,甚至超过人类识别准确率的两个数量级,引起了科研工作者的广泛关注[5]. 同时,卷积神经网络的权值共享网络结构使之更类似于生物神经网络[4],降低了网络模型的复杂度,减少了权值的数量. 由于这种网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性,近年来在图像处理中得到了广泛的使用: 2012年,Krizhevsky等[6]使用卷积神经网络搭建的系统,在ImageNet图像数据集中将分类错误率从25%下降到17%;2014年,Facebook搭建的卷积神经网络系统在人脸验证上将正确率提高到97.25%(人眼辨识的正确率是97.53%)[7]. 由于卷积神经网络卷积神经网络在计算机视觉、图像处理中成功应用,近两年来研究者们开始将其应用到语音识别领域. 2012年多伦多大学初步建立了卷积神经网络用于语音识别的模型结构,并同深层神经网络相比取得相对10%的性能提升[8]. 随后IBM和Microsoft也都与多伦多大学合作在2013年发表了相关文章,验证了卷积神经网络相对深层神经网络建模的有效性[9-10].

与深层神经网络相比,卷积神经网络的关键在于引入了卷积和聚合(又作采样)的概念. 卷积神经网络通过卷积实现对语音特征局部信息的抽取,再通过聚合加强模型对特征的鲁棒性. 本文深入分析了卷积神经网络中卷积层和聚合层的不同结构对识别性能的影响情况,并与目前广泛使用的深层神经网络模型进行了对比. 相比深层神经网络,卷积神经网络能够在保证识别性能的同时,大幅度降低模型的复杂度(规模). 同时,卷积神经网络也具有更合理的物理意义,由此降低对前段语音特征提取的依赖. 本研究在标准英文连续语音识别库TIMIT[11]以及汉语电话自然口语对话数据集上面进行了实验,对卷积神经网络的输入特征、卷积器尺寸和个数、计算量和模型规模等做了详细的对比实验.

1 卷积神经网络

卷积神经网络由一组或多组卷积层+聚合层构成[4]. 一个卷积层中包含若干个不同的卷积器,这些卷积器对语音的各个局部特征进行观察. 聚合层通过对卷积层的输出结点做固定窗长的聚合,减少下一层的输入结点数,从而控制模型的复杂度. 一般聚合层采用最大聚合算法(max pooling),即对固定窗长内的结点选取最大值进行输出. 最后,通过全网络层将聚合层输出值综合起来,得到最终的分类判决结果. 这种结构在图像处理中获得了较优的性能[12]. 卷积神经网络相比深层神经网络等神经网络结构,引入三个重要的概念——局部卷积、聚合和权值共享[5].

图1给出了卷积神经网络用于语音识别声学建模时,典型的卷积层和聚合层的结构. 当二维图像作为卷积神经网络的输入时,两个维度上特征的物理意义是完全一样的. 将语音看作二维特征输入时,第一维是时域维度,第二维是频域维度,这两维的物理意义完全不同. 由于深层神经网络上实验证明,多帧串联的长时特征对模型性能的提高非常重要,在卷积神经网络的输入特征上,也保留了该方法,将当前帧的前后几帧串联起来构成长时特征. 考虑到差分特征对静态特征的补充关系,实验中将差分特征一起串联在长时特征中,这样构成的特征作为卷积神经网络的第一维特征. 在卷积神经网络的另一维——频域维度上,一般采用梅尔域的滤波带系数(filterbank)作为参数(如图1中选择N个滤波频带). 卷积神经网络中卷积层的物理意义可以看做,通过卷积器对局部频域的特征观察,抽取出局部的有用信息(局部卷积). 这里,将同一种卷积器作用在不同的滤波带上,每个滤波带包含有当前帧该滤波带的系数,以及该滤波带上的长时特征,通过下式计算得到卷积器的输出:

$$C_{i,k}=\theta\left(\sum_{b=1}^{s-1}\boldsymbol{w}_{b,k}\boldsymbol{v}_{b+i}^{\mathrm{T}}+a_k\right). \tag{1}$$

式中,$\boldsymbol{v}_{b+i}^{\mathrm{T}}$为第$i$组输入特征矢量,$\boldsymbol{w}_{b,k}$为第$k$个卷积器的权值参数,$s$为卷积器的宽度,$a_k$为网络偏置. 通过将第$i$组输入和第$k$个卷积器做加权平均后,通过非线性函数$\theta$得到卷积层的一个输出结点值,$\theta$一般选择反正切函数或sigmoid函数.

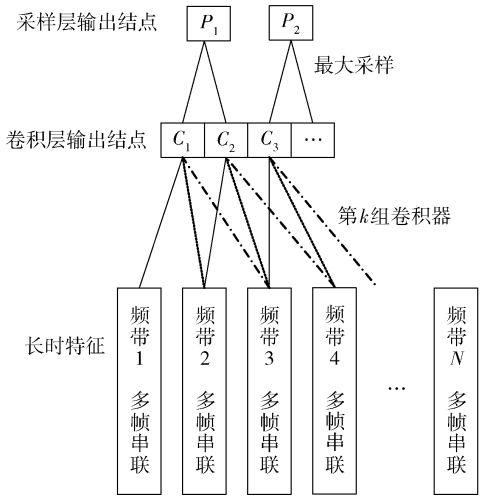

图1 卷积神经网络中卷积层和最大聚合(采样)层的示例图

Fig. 1 Diagram of the CNN convolution layer and max-pooling layers

由此得到的输出为该种卷积器对局部特征的观察结果. 由于使用的是相同的卷积器，其卷积参数完全相同，存储时只需保留一组卷积参数(权值共享). 另一方面，由于一种卷积器所能观察的信息有限，所以一般会使用多种不同的卷积器从不同视角上进行观察，从而得到更多的信息量. 最终的存储量仅为各种卷积器的自由参数量之和，相比深层神经网络全网络连接结构，大大减少了模型的存储规模. 同时，卷积运算的一个重要特点就是，通过卷积运算，可以使原信号特征增强，并且降低噪音，这也使得基于卷积运算的卷积神经网络模型有着更好的抗噪性能.

在卷积层之后，紧跟着的是聚合层. 在语音识别中，采用最大聚合算法(聚合). 以图 1 为例，从 C_1 和 C_2 这两个卷积层输出结点中选择最大值作为聚合层的输出 P_1. 这样做的好处: 一是可以减少输出结点数，控制模型的计算量; 二是通过对几个结点选最大值进行输出，增加模型对语音特征的鲁棒性.

到目前为止，卷积神经网络的信息都还是停留在局部观察的结果. 要得到最终的分类结果，需要将这些信息综合起来. 所以在卷积层之后，通过一个全网络层，将聚合层的各个输出综合起来，最后通过输出层得到各个状态的分类后验概率.

2　实验结果和分析

2.1　实验条件

核心对比实验在英文标准连续语音识别库 TIMIT 上进行，性能指标有神经网络的验证集分类帧正确率(frame correct rate)和最终的音素识别正确率(phone correct rate). 使用 462 个说话人的语音作为训练集，另外 144 个说话人的语音作为神经网络的验证集. TIMIT 提供的 24 人的 core 测试集作为测试集. 各个集之间无说话人重叠. 在特征提取部分，使用传统的 25 ms帧长、10 ms 帧移的方式提取特征. 40 维的梅尔域滤波带系数作为特征输入，同时也包含其一阶和二阶差分系数. 在送入卷积神经网络训练前，将多帧串联构成长时特征. 所有特征都进行了逐句的均值方差规整. 实验中采用各 5 帧，总 11 帧的串联长时特征.

卷积神经网络的训练采用一层卷积层 + 聚合层和一层全网络层的结构. 为了与之对比，训练了深层神经网络模型，采用的是两个隐含层结构，保证和卷积神经网络的网络层数一致. 卷积神经网络和深层神经网络的目标分类都为 183 个音素状态(61 个音素，每个音素三个状态)，其输出层为该帧属于某个音素的后验概率，通过贝叶斯公式将其转化成似然概率应用于解码阶段. 在实验中，为了直接观察卷积神经网络在声学建模上的性能，采用了不带语言模型的音素解码.

2.2　实验结果

2.2.1　卷积神经网络和深层神经网络对比

表 1 给出了卷积神经网络和深层神经网络在不同条件下的性能对比结果. 在特征方面，尝试了不使用和使用一阶和二阶差分特征两种方式，分别对应表中的“40 维特征”和“120 维特征”. 卷积神经网络的结构为两个隐层: 第一个隐层为卷积层 + 聚合层，卷积器种类为 100 种，对应两种不同特征时卷积器参数分别为 11×8 和 33×8，聚合层为 3 个结点选择一个最大输出的方式; 之后紧接一个 1024 结点的全网络隐层. 基于 120 维特征输入的结构，卷积神经网络的总模型大小为 2.6×10^6，总计算量为 1.6×10^6 次(矩阵乘法). 深层神经网络也为两个隐层，每层都为 1024 结点的全网络连接. 同样基于 120 维特征输入的结构，深层神经网络的总模型大小为 10.1×10^6，总计算量为 2.6×10^6 次(矩阵乘法). 对比看到，无论是模型规模还是实际计算量，卷积神经网络都比深层神经网络更小. 在这样的条件下，表 1 结果显示无论是选择不使用或使用一阶和二阶差分特征，卷积神经网络的帧正确率和音素正确率都稳定优于深层神经网络; 并且，使用一阶和二阶差分特征会进一步提高模型性能.

表 1　TIMIT 测试集上卷积神经网络和深层神经网络的参数和性能对比

Table 1　Performance comparisons between CNN and DNN on TIMIT corpus

模型	输入维数	卷积器个数	卷积器形状	聚合层	全连接隐层	帧正确率/%	音素正确率/%
卷积神经网络	40	100	11×8	1×3	1024	47.6	61.7
卷积神经网络	120	100	33×8	1×3	1024	53.6	66.3
深层神经网络	40	—	—	—	1024×1024	46.3	60.1
深层神经网络	120	—	—	—	1024×1024	51.8	64.6

在接下来的实验中，基于 120 维特征(含一阶和二阶差分特征)输入，对卷积神经网络的卷积层和聚合层进行了不同参数条件下的性能对比. 不同参数所对应的物理意义不相同，通过实验寻找到用卷积神经网络表征语音的最优方式.

2.2.2　卷积层

卷积神经网络通过卷积器对局部特征进行分析，通过聚合层加强抽取出来的特征鲁棒性，最后通过全

网络层建立模型得到最后的分类结果. 在这个过程中,由于卷积器肩负了直接对输入原始特征的分析、抽取过程,使得卷积器的设计成为卷积神经网络的重点. 卷积器的参数有两个: 卷积器个数和卷积器形状. 下面分别从这两个方面分析不同参数对分类性能的最终影响.

表 2 为在 TIMIT 测试集上不同卷积器个数的卷积神经网络性能对比,分别给出了各个模型的验证集帧正确率和测试集的音素正确率. 在这组实验中,除了卷积器个数不同以外,所有卷积神经网络的其他参数都保持一致: 卷积器形状 33×5,聚合层 1×4,全连接单隐层 1024,输出分类 183 类. 随着卷积器个数从 50 个逐步上升到 200 个,帧正确率和音素正确率都有稳步的提升,特别是当卷积器从 50 个上升到 100 个时,性能有超过 1% 的提高,再继续增加卷积器个数到 200 个时,性能基本没有变化. 实验现象表明: 不同卷积器可以从不同的角度提取出不同的信息,如果个数太少,则会导致提取的信息量受限,卷积神经网络的建模性能也就受到影响,所以在卷积神经网络中,要想更好地表征语音特性,卷积器的个数不能太少(同时也不宜太多. 太多会增加计算量,并且性能已经基本饱和).

表 2 TIMIT 测试集上不同卷积器个数的卷积神经网络性能对比

Table 2 Performance comparisons between CNNs with different numbers of convolution filters on TIMIT corpus

卷积器个数	帧正确率/%	音素正确率/%
50	53.1	65.7
100	53.9	66.8
150	54.1	67.1
200	54.3	67.1

表 3 为在 TIMIT 测试集上不同卷积器形状的卷积神经网络性能对比. 卷积器的形状主要是指对局部多大范围的特征进行观察,理论上观察得越细越有可能发现局部的有用信息,但同时也可能会牺牲模型的泛化能力,使得在识别非匹配语音时效果变差. 在这组实验中,除了卷积器形状不同以外,所有卷积神经网络的其他参数都保持一致: 卷积器个数 100,聚合层 1×3,全连接单隐层 1024,输出分类 183 类. 表中对比了 $33 \times A$ 的卷积器形状对性能的最终影响. 之所以固定卷积器第一维参数为 33,是考虑到输入的特征是 11 帧串联,包含有 0、1 和 2 共三阶差分信息,这样构成了 33 维参数. 真正需要卷积器细节观察的应该是不同频带上的特征分布. 实验中选择 40 个频带的特征输入,当卷积器的第二维参数为 A 时,则表示这 40 个频带上每连续 A 个频带作为一个观察窗,送入卷积器抽取出相应信息(频带窗移为 1 个频带的长度,也就是相邻两个观察窗有 $A-1$ 的长度的频带交叠(overlapping)).

从结果中看到,虽然随着卷积器形状的逐步细化,音素正确率有所提高,但是幅度微弱. 这说明卷积器的形状对性能的影响相对不明显,同时为了得到更好的泛化性,一般选择比较适中的卷积器形状.

表 3 TIMIT 测试集上不同卷积器参数的卷积神经网络性能对比

Table 3 Performance comparisons between CNNs with different convolution filter sizes on TIMIT corpus

卷积器形状	帧正确率/%	音素正确率/%
33×8	53.6	66.3
33×5	53.8	66.7
33×2	53.8	66.9

2.2.3 聚合层

除卷积层以外,聚合层也是卷积神经网络结构的特点之一,聚合是为了加强模型的鲁棒性. 通常语音识别系统的性能会在不同环境、不同说话人等情况下受到影响,主要是由于不同的环境或说话人会使得频谱特征发生偏移. 由于聚合本身是对相邻几个观察窗的输出做最大值选择,相当于模糊语音特征,即使发生偏移,也不影响最大值的选择,从而加强模型的鲁棒性.

表 4 给出了卷积器形状为 33×5 时不同聚合参数下卷积神经网络性能. 聚合层为 1×1,表示不进行聚合,每个观察窗的输出都将作为下层的输入送入训练. 相比表中的 $1 \times M$ ($M > 1$) 的聚合结构,不聚合的最终识别性能明显变差($> 2\%$),这个结果充分说明聚合对性能保证的必要性. 在使用聚合的结构中,M 的选择对最终的音素识别率影响非常微弱. 考虑到 M 越大,聚合层输出结点数就越少,全网络层的计算量和规模也越小. 为了有效控制模型的规模,实际中一般 M 不会选的太小.

表 4 TIMIT 测试集上不同聚合参数的卷积神经网络性能对比

Table 4 Performance comparisons between CNNs with different max-pooling structures on TIMIT corpus

聚合层	帧正确率/%	音素正确率/%
1×1	52.3	64.1
1×3	53.8	66.7
1×4	53.9	66.8
1×6	53.7	66.6

2.2.4 泛化性

之前的实验都是基于单(卷积层 + 聚合层) + 单(全网络隐层)的结构. 由于神经网络的性能一般与训练数据量和模型规模成正比,所以试加大卷积神经网络的网络层数,并与同等层数的深层神经网络模型性能进行对比. 选择两种典型的卷积神经网络结构: 两个(卷积层 + 聚合层) + 单(全网络隐层)的结构; 单

（卷积层 + 聚合层）+ 两个（全网络隐层）的结构. 与之对比的是含有三个全网络隐层的深层神经网络模型. 表 5 给出了这三种模型的性能. 在卷积神经网络 1 中，第二卷积层使用了比第一层更多的卷积器数目，这样做借鉴了图像处理中逐步细化的思想：空间分辨率递减，每层所含的卷积器数递增，这样可用于检测更多的特征信息. 卷积神经网络 2 中仍只含单（卷积层 + 聚合层），但通过增加一个全网络隐层达到相同的网络层数. 卷积神经网络 1 和卷积神经网络 2 与全网络层相连的聚合层输出结点数都基本保持在 1000 左右. 深层神经网络则是通过直接训练三个隐层保持与卷积神经网络一致的模型层数.

首先，对比深层神经网络和卷积神经网络，无论是帧正确率还是音素正确率，深层神经网络的性能都不如卷积神经网络，这与两个隐层的结论一致，也再次证明卷积神经网络结构的性能更优. 其次，对比卷积神经网络 1 和卷积神经网络 2，发现卷积神经网络 2 的帧正确率比卷积神经网络 1 的更高，但最后音素正确率上卷积神经网络 1 却表现最好. 卷积神经网络 1 和卷积神经网络 2 的最大不同在于，卷积神经网络 1 用一个（卷积层 + 聚合层）替代了卷积神经网络 2 的一个全网络层. 实际上，（卷积层 + 聚合层）的结构大大减少了模型可训练的参数量，这样可以有效避免模型过分模拟到训练数据上，导致过训练. 从这点来看，（卷积层 + 聚合层）的结构有着比全网络连接结构更强的泛化能力.

表 5　TIMIT 测试集上三个隐层的卷积神经网络和深层神经网络性能对比

Table 5　Performance comparisons between CNN and DNN with three hidden layers on TIMIT corpus

模型	设置项目	卷积器个数	卷积器形状	Max-pooling	全连接隐层	帧正确率/%	音素正确率/%
卷积神经网络 1	第一层卷积	100	33 × 8	1 × 3	1024	53.7	66.9
	第二层卷积	200	1 × 2	1 × 2			
卷积神经网络 2	第一层卷积	200	33 × 8	1 × 3	1024 × 1024	54.0	66.3
深层神经网络	120 维输入	—	—	—	1024 × 1024 × 1024	52.3	65.1

2.2.5　大规模数据集实验

上述 TIMIT 数据集的实验对比结果显示卷积神经网络相对深层神经网络的建模性能更好. 考虑到 TIMIT 数据库的数据总量较小，为了使研究结果具有一定的推广性，进一步在大规模数据集上进行对比. 实验在汉语普通话大词表非特定人电话自然口语对话系统中进行. 识别指标为汉字字错误率 WER（word error rate）. 所有数据集皆为通用标准数据集，包含训练集和测试集两部分. 其中，训练数据来自语言数据联盟 LDC 提供的汉语普通话数据：Call—Home、Call—Friend 以及 Call—HKUST，总共 100 h. 测试数据 HDev04 是由香港大学 2004 年采集的电话自然口语对话数据，它包含了 24 个电话对话，全集长度大约 4 h. 为缩短实验周期而不失其统计意义，从中随机选取出了 1 h 的数据用于测试[13]. 测试数据 86305 是国家 863 组织的语音识别评测在 2005 年的测试集[14].

识别系统中采用汉语普通话通用发音字典，共 43514 个词条. 语言模型采用三元文法语言模型，语料来源除训练集标注文本外，还包含华盛顿大学公布的汉语普通话文本语料，2005 年 863 语音识别电话语音评测数据的训练文本，以训练文本常用词为关键词利用 google 搜索的文本语料，以及其他部分自行下载的文本语料. 详细说明可参见文献[14].

在神经网络建模方面，网络输入特征采用 39 维的 filterbank 特征，做三阶扩展到 117 维，串联 13 帧数据作为输入，输出为 6245 个状态. 深层神经网络和卷积神经网络都为三个隐层的结构. 深层神经网络的输入维数为 1521 维，隐层每层结点数为 1024. 卷积神经网络的输入端为 39 × 39 的结构，隐层为两个卷积层 + 一个全连接层的结构. 第一个卷积层采用 7 × 7 的滤波器，总共 64 种滤波器，聚合层为 3 × 3 结构；第二个卷积层采用 4 × 4 的滤波器，总共 128 种滤波器，聚合层为 2 × 2 结构. 表 6 给出了深层神经网络和卷积神经网络的字错误率对比结果：卷积神经网络相对于深层神经网络字错误率有 1% 的下降. 实验结果证明无论是在小规模数据集还是大规模数据集上，卷积神经网络的性能都一致优于深层神经网络.

表 6　汉语普通话大词表非特定人电话自然口语对话系统中深层神经网络和卷积神经网络的对比结果

Table 6　Performance comparisons between DNN and CNN for conversational LVCSR task

神经网络	测试集	WER/%
深层神经网络	86305	48.1
卷积神经网络	86305	47.1
深层神经网络	HDev04	51.6
卷积神经网络	HDev04	50.7

3　结论

对比了卷积神经网络中卷积层和聚合层的不同结构对识别性能的影响情况，对卷积神经网络的输入特

征、卷积器尺寸和个数、计算量和模型规模等做了详细的对比实验,并与普遍使用的深层神经网络进行了对比. 卷积神经网络通过卷积层对局部特征进行观察,再经过全网络层的信息整合最终得到输出概率,相比深层神经网络具有更好的物理意义. 同时,由于卷积神经网络的权值共享,使得模型复杂度大大降低. 在多个标准库上的实验证明,在计算量比深层神经网络更少的条件下,卷积神经网络的识别性能更优,泛化能力更强.

参 考 文 献

[1] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans Audio Speech Lang Process*, 2012, 20(1): 30

[2] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*, 2012, 29(6): 82

[3] Yu D, Deng L. Deep learning and its applications to signal and information processing. *IEEE Signal Process Mag*, 2011, 28(1): 145

[4] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series // *The Handbook of Brain Theory and Neural Networks*, 1995

[5] Fan B L. *Research on Parallelization of Convolutional Neural Networks* [Dissertation]. Zhengzhou: Zhengzhou University, 2013
(凡保磊. 卷积神经网络的并行化研究[学位论文]. 郑州,郑州大学, 2013)

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks // *Advances in Neural Information Processing Systems*, 2012: 1097

[7] Wolf L. DeepFace: closing the gap to human-level performance in face verification // *IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, 2014

[8] Abdel-Hamid O, Mohamed A, Jiang H, Penn G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition // 2012 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Kyoto, 2012: 4277

[9] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR // 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Vancouver, 2013: 8614

[10] Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition // *INTERSPEECH*. Lyon, 2013: 3366

[11] TIMIT. *Linguistic Data Consortium* [DB/OL] [2014-08-10]. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[12] LeCun Y, Huang F J, Bottou L. Learning methods for generic object recognition with invariance to pose and lighting // *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, 2004: II-97-104

[13] Zhang Q Q, Pan J L, Yan Y H. Tonal articulatory feature for Mandarin and its application to conversational LVCSR // *Tenth Annual Conference of the International Speech Communication Association*. Brighton, 2009: 3007

[14] Zhang Q Q, Cai S, Pan J L, et al. Improved acoustic models for conversational telephone speech recognition // 9*th International Conference on Fuzzy Systems and Knowledge Discovery* (*FSKD*). IEEE, 2012: 1229