深度学习在语音识别中的研究进展综述*

侯一民[1]，周慧琼[1†]，王政一[2]
(1. 东北电力大学 自动化工程学院，吉林 吉林 132012；2. 中国航空规划设计研究总院有限公司，北京 100120)

摘 要：在当今的大数据时代里，对于处理大量未经标注的原始语音数据的传统机器学习算法，很多都已不再适用。与此同时，深度学习模型凭借其对海量数据的强大建模能力，能够直接对未标注数据进行处理，成为当前语音识别领域的一个研究热点。主要分析和总结了当前几种具有代表性的深度学习模型，介绍了其在语音识别中对于语音特征提取及声学建模中的应用，最后总结了当前所面临的问题和发展方向。

关键词：机器学习；深度学习；语音数据；语音识别
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2017)08-2241-06
doi:10.3969/j.issn.1001-3695.2017.08.001

Overview of speech recognition based on deep learning

Hou Yimin[1], Zhou Huiqiong[1†], Wang Zhengyi[2]
(1. *School of Automation Engineering, Northeast Dianli University, Jilin Jilin* 132012, *China*; 2. *China Aviation Planning & Design Institute Co. Ltd., Beijing* 100120, *China*)

Abstract: In the era of big data, many of traditional machine learning methods of disposing unlabeled raw voice data have become less applicable. At the same time, deep learning models can directly process unlabeled data because of its powerful capability of modeling to deal with the massive data, and has become a hot research in the field of speech recognition. To begin with, this paper analyzed and summarized the state-of-the-art deep learning of models. And then, it discussed the applications to speech recognition with speech features extraction and acoustic modeling. Finally, it concluded the problems faced and development orientation.

Key words: machine learning; deep learning; voice data; speech recognition

随着移动互联网的不断发展，实现人与计算机之间的自由交互越来越受到人们的重视。用语音来实现这一目标，主要包括三项技术，即语音识别、语音编码和语音合成[1]。本文所研究的自动语音识别(automatic speech recognition, ASR)技术，主要是完成语音到文字的转变[2]，属于非特定人语音识别。语音识别发展到现在，已经改变了人们生活的很多方面，从语音打字机、数据库检索到特定的环境所需的语音命令，给人们的生活带来了很多方便。对于语音识别系统，最具有代表性的识别方法有特征参数匹配法、隐马尔可夫法和神经网络法[1]。对于神经网络，2006 年以前，人们尝试训练深度架构都失败了，用浅层网络的学习训练一个深度的有监督前馈神经网络是失败的，失败的主要原因是梯度不稳定，并且监督学习数据的获取也非常昂贵，梯度下降算法对初始值的敏感也使深度网络参数难以训练，最后还是将其变为浅层(只包含 1 ~ 2 个隐层)。直到 2006 年，Hinton 等人[3]提出逐层贪婪无监督预训练深度网络之后，微软成功地将深度学习应用到自己的语音识别系统中，比起之前的最优方法，使单词错误率降低了约 30%[4]，这称得上是语音识别领域中的再一次重大突破。随后，微软的基于上下文相关的深度神经网络—隐马尔可夫模型(context-dependent DNN-HMM, CD-DNN-HMM)对大词汇量语音识别的研究成果，彻底改变了语音识别系统的原有技术框架[5]。目前许多国内外知名研究机构，如微软、讯飞、Google、IBM 都积极开展对深度学习的研究[6]。在人们生活的应用层面上，由于移动设备对语音识别的需求与日俱增，以语音为主的移动终端应用不断融入人们的日常生活中，如国际市场上有苹果公司的 Siri、微软的 Cortana 等虚拟语音助手；国内有百度语音、科大讯飞等。还有语音搜索(VS)、短信听写(SMD)等语音应用都采用了最新的语音识别技术。现在，绝大多数的 SMD 系统的识别准确率都超过了 90%，甚至有些超过了 95%，这意味着新一轮的语音研究热潮正在不断兴起。

1 深度学习神经网络

1.1 现有机器学习的局限性

深度学习(deep learning)是机器学习研究领域中的一个分支，可理解为人工神经网络的发展[7]，本质上是训练深层结构模型的方法，也是对于通过多层来表示对数据之间的复杂关系进行建模的算法。深层结构是相对于浅层结构而言的[8]，当前回归、分类等多数学习方法一般都是浅层结构的算法，在有限的有标签样本和大量无标签样本的情况下，少量计算单元的表示能力有限，进而其泛化能力也受到了一定的制约[9,10]。并且浅层模型还有一个重要特点，就是需靠人工经验来抽取样本的特征，它强调模型的主要任务是分类或预测。则在模型不变

收稿日期：2016-09-19；**修回日期**：2016-11-22 **基金项目**：国家自然科学基金资助项目(61403075)；吉林省科技发展计划资助项目(20150414051GH)

作者简介：侯一民(1978-)，男，副院长，博士，主要研究方向为模式识别与智能系统、检测技术与自动化装置等；周慧琼(1992-)，女(通信作者)，硕士研究生，主要研究方向为深度学习、语音识别(342277258@qq.com)；王政一(1991-)，男，硕士，主要研究方向为智能控制.

的情况下,特征选取的好坏就成为了整个系统性能的关键部分,这通常需要大量的人力去发掘更好的特征,而且也需要大量时间去调节,很费时费力[7]。

1.2 深度学习主要思想

在2006年之前,很多人尝试用传统训练算法(如BP算法)去训练深度架构都以失败告终,最后还是将其变为浅层(2至3层),虽被称做多层感知器(multi-layer perception,MLP),但实际上还是浅层结构模型。由Hinton等人发表的具有革命性的深度信念网(deep belief network,DBNs)引领着后人对深度学习(deep learning)的研究[3,11,12],其主要思想包括:

a)从底往上的非监督学习,就是用无标签数据进行每一层的预训练(pre-training),而每一层的训练结果作为其高一层的输入,这是与传统神经网络相比最大的区别,这个过程可看做是特征学习(feature learning)的过程。

b)从顶向下的监督学习,就是用有标签的数据调整所有层的权值和阈值,按照误差反向传播算法(back propagation,BP)自顶向下传输,对网络进行微调(tune-fining)。

由于深度学习的第一步不是跟传统神经网络一样去随机初始化,而是通过学习数据的结构得到,所以这个初值更接近全局最优,进而取得更好的结果。因此相比BP算法,深度学习算法效果好,要归功于第一步的特征学习。

深度学习网络就像人类大脑的学习机制一样,在面临大量的感知信息时,通过低层特征的组合形成更加抽象的高层特征,学习到数据的分布式特征,从而可像人脑一样实现对输入信息的分级表达来表示信息的属性或类别[13,14]。并且,此方式相比浅层结构的机器学习,提取的特征是靠网络自动完成的,不需要人工参与[15]。同样是人工神经网络,深度学习神经网络(DNN)对于传统的神经网络,突破之处是它的网络层数和解决深层网络训练难度的方法[16]。也可以认为DNN能在任意隐层分开,其下面的所有隐层被认为是特征变换,其上所有层可认为是分类模型。

1.3 深度模型介绍

1.3.1 限制波尔兹曼机

限制波尔兹曼机(restricted Boltzmann machine,RBM)[17]属于无监督模型,它的子模块有两层,每层中的各节点之间是没有连接的,属于Markov随机场的一种特殊情况[18]。第一层为可视层,第二层为隐藏层。一个RBM中包含权值、可视层偏置、隐藏层偏置这三个模型参数。

根据不同类型的节点,能量模型和相对应的条件概率也将不同[19]。有伯努利(可视)—伯努利(隐藏)(Bernoulli-Bernoulli)RBM模型和高斯(可视)—伯努利(隐藏)(Gaussian-Bernoulli)RBM模型。在DBN中,最基本的RBM是所有节点为随机二值变量节点(取0或1)。但值得注意的是,一般语音数据在DBN中作为输入数据,输入层节点状态须满足高斯分布,所以采用Gaussian-Bernoulli RBM模型,可将随机的实数变量转换到随机的二进制变量,然后进一步利用Bernoulli-Bernoulli RBM模型或Bernoulli-Gaussian RBM模型处理,网络最后两层一般为Bernoulli-Gaussian RBM模型,将语音的二值变量数据转换为实数变量易于研究,这种DBN一般用于语音信号的深层特征提取[20]。

由多个RBM组成的神经网络称为深度信念网络[19]。最近,研究者对于DNN与DBN进行了更加细致的区分,将这种使用DBN去预训练DNN初始参数的网络称为DBN-DNN[5,6]。它是先用RBM进行非监督学习确定网络初始值(降低了对数据的要求,一般仅到这一步的无监督训练,可用于提取语音深层特征),然后结合输出层用少量的有标签数据自顶向下地使用BP算法对网络进行微调,所以DBN-DNN是一种有监督和无监督相结合的混合模型。这种混合模型与文献[20]提到的有监督深度模型类似,在ASR中既可以提取特征,也可以取代HMM声学模型,直接进行语音识别。

1.3.2 自动编码器

自动编码器(auto encoder,AE)是由自动关联器(auto associator)[21]演变而来的。自动关联器是一种MLP结构,其中输出、输入维度一样,并定义输出等于输入。为了能够在输出层重新产生输入,MLP需找出输入在隐藏层的最佳表示。一旦训练完成,从输入到隐藏层的第一层充当编码器,而隐藏层单元的值形成编码表示;从隐藏单元到输出单元的第二层充当解码器,由原信号的编码表示重构原信号。

因为输入是无标签数据,AE利用自动关联器的这一点,将重构原信号与原信号之间的差作为目标函数,以调整编码器(encoder)和解码器(decoder)的参数,使这个重构误差最小;调整完后,从隐层单元到输出单元的解码器权值参数就不需要了(被去掉),直接将隐层单元的值,也就是编码值,作为第二个自动编码器的输入,训练方式与之前一样。最后可在自动编码器的最顶编码层添加一个分类器[23],通过有标签样本进行监督训练方法(梯度下降法)对网络进行微调,这跟DBN-DNN类似。

由多个自动编码器堆叠而成的网络称为深度自动编码器(deep autoencoder,DAE)[22],它属于无监督模型。而自动编码器还可变形为去噪自动编码器(denoising autoencoder)[23]和稀疏编码(sparse coding)。还可将前文提到的RBM节点模型运用到自动编码器神经网络的节点训练当中[20],构成的DNN对ASR系统中的语音特征MFCC进行可监督和无监督的提取。

1.3.3 卷积神经网络

对于语音识别早期发展起来的延时神经网络(time-delay neural network,TDNN)可以看成是卷积神经网络(convolutional neural network,CNN)[24]的一种特殊情况或前身,即共享权值被限制在单一的时间维度上,且没有池化层,适用于语音和时间序列信号的处理。但最近研究人员发现,在语音识别领域中,时间维度上的不变形并没有频率维度上那么重要[25,26],所以将CNN用于语音识别中是再适合不过了。CNN是第一个真正多层神经网络结构学习的算法,因为在没出现无监督的逐层贪婪预训练深层神经网络之前,用BP算法训练深度神经网络几乎不能实现,但卷积神经网络就是一个特例,它是利用减少空间关系参数的数目以提高一般前向BP训练速度[27]。

1.3.4 递归神经网络

无论是CNN还是DBN或SAE,都没有考虑样本之间的关联性,而在递归(循环)神经网络(recurrent neural network,RNN)[21]中,除了前馈连接之外,单元具有自连接或到前面层的连接,这种递归性充当短期记忆,并使得网络记得过去发生的事。它可以认为是另一类用于无监督(和有监督)学习的深度网络。在大部分情况下使用部分递归网络,其中有限个递归连接被添加到多层感知器中,这结合了多层感知器的非线性逼近能力和递归的时间表达能力的优点。与权值共享一样,方法

是按时间对不同权重的改变求和,并用平均值更新权重。在无监督学习模式下,RNN 的预训练可使用 RBM 或 AE 来进行网络参数初始化[28]。RNN 被用来根据先前的数据样本预测未来的数据序列,并且学习过程中没有用到类别信息,非常适合序列数据(如语音、文本)建模,但由于在训练中遇到的梯度弥散或梯度爆炸问题,RNN 很难训练来捕捉长时相关性。最近在 Hessian-free 优化研究方面的进展,在一定程度上解决了这个问题,该方法使用了近似二阶信息或随机曲率估计。而 Bengio 和 Sutskever[29,30] 探索了不同的用于训练生成式 RNN 的随机梯度下降优化算法,并证明了这些算法超越 Hessian-free 优化方法。

2　基于深度学习的语音识别发展及研究现状

如图 1 所示,对输入的训练语音进行预处理、提取特征参数并训练声学模型;而语言模型是通过从训练语料(通常是文本形式)学习词或句之间的相互关系来估计假设词序列的可能性,又叫做语言模型分数;解码搜索是对测试语音经过预处理和特征提取后的特征向量序列与若干假设词序列计算声学模型分数与语言模型分数,最后将总体输出分数最高的词序列当做识别结果[31~33]。

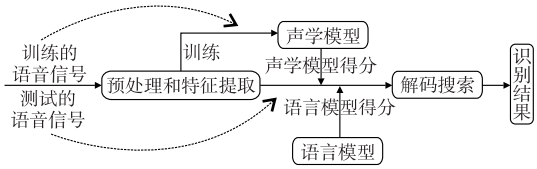

图1　语音识别过程的总体结构

虽然人工神经网络已经存在了大半个世纪,但深层结构的神经网络对语音识别产生重要影响还是从 2010 年开始的[34,35]。在 2013 年的声学、语音和信号处理国际会议(ICASSP)上讨论了语音识别和相关应用的深度神经网络学习的新类型;还有在 2011 年、2013 年的机器学习国际会议(ICML)上讨论了关于音频、语音和视觉信息处理的学习结构、表示和最优化[36]。目前深度学习理论已成功应用于孤立词识别、音素识别、声韵母识别和大词汇量连续语音识别(LVCSR)[37~40],其应用主要集中于提取更具有表征能力的语音数据高层特征以及加强对现有的基于 HMM 的声学模型的构建[41]。

2.1　利用深度学习进行语音特征提取

人们进行了许多研究致力于研究模仿人类听觉过程的特征,像 MFCC 特征,而本文更强调的是对于 DNN 能学习到深层语音特征的这一重要发展。

在过去,最流行的语音识别系统通常使用 Mel 倒谱系数(Mel frequency cepstral coefficient, MFCC)[42] 或者相对频谱变换—感知线性预测(perceptual linear prediction, RASTA-PLP)[43] 作为特征向量,使用高斯混合模型—隐马尔可夫模型(Gaussian mixture model-HMM, GMM-HMM)作为声学模型,用最大似然准则(maximum likelihood, ML)和期望最大化算法来训练这些模型。利用传统的语音特征提取算法(如 MFCC 或 PLP)提取的特征只对单帧信号作用,不能很好地涵盖有效语音信息,也易受噪声污染。对于语音的特征学习和语音识别而言,这个目标可以归纳为对原始频谱特征的使用或是对波形特征的使用。过去 30 年以来,虽然对语音频谱进行变换丢失了原始语音数据的部分信息,但是多种“手工制作”的特征促进了 GMM-HMM 系统识别率的巨大提升。其中最成功的是非自适应的余弦变换,它促进了 MFCC 特征的产生。余弦变换近似地去除了特征成分之间的相关性,这对使用对角协方差阵的 GMM 来说很重要。然而,当深度学习模型(DNN、DBN)、深度自编码器替代 GMM 后,使得去除特征之间的相关性变得无关紧要。

研究人员用 DNN 来提取语音特征的模型中,就有用到 DBN。在文献[44]中提出利用 DBN 来进行音素识别,与传统语音识别系统性能相比进行有较高提升,DNN 在其中充当一个单独的特征提取器,为传统的 GMM-HMM 提供特征。随后,研究人员将 DBN 进行改进,采用瓶颈深度信念网络(bottleneck deep belief network, BN-DBN)来提取语音特征[45]。BN-DBN 一般设定网络层数为奇数,而将其中间一层定义为瓶颈层。瓶颈层的神经元个数与单帧的维数相等,便于接下来训练声学建模。实验结果也证明了 BN-DBN 提取的语音特征表现要好于现有的 MFCC 等其他特征。文献[46]中提出了一种新的深层瓶颈特征类型——通过剪枝节点方法重构 DNN。经过节点修剪之后的网络拓扑结构减少了冗余,从而得到新的派生 DNN 特征,其对干净语音的最优识别错误率仅为 7.3%,而对带噪语音识别错误率为 23.8%。随后,在文献[47]中证明了经过这种节点修剪及重构方法处理的 DNN 比原有的网络结构减少了 85%,使训练速度提高了 4.2 倍。文献[48]中提到,将 DBN 作为 GMM-UBM 说话人确认系统中的特征提取器,实验结果显示当采用四个隐层的 DBN 进行测试时识别错误率仅为 9.75%。以上实验一般都是只将 DNN 学习到的特征作为 GMM-HMM 系统的输入,但文献[49]指出,使用 MFCC 和 DNN(BN-DBN)后验结合的特征在低噪或中等噪声情况下的串联结构,更优于只用 DNN 特征。

另一种采用 DAE 神经网络来进行语音特征提取,是继 DBN 后提出的。文献[20]利用 DAE 在 863 汉语语音库中进行了实验,结果证明了经过 DAE 的无监督模型或由 DAE 改编的深度有监督模型提取得到的语音特征都要好于传统 MFCC 特征。需注意的是,此文献中无监督模型是将多个隐层的最中间层的输出特征作为 GMM-HMM 识别器的输入;而有监督模型提取的特征则是将网络最后一层隐层的输出特征作为识别器的输入。由于 CNN 在计算机视觉、图像处理中的成功应用,近两年来研究人员开始将其应用到语音识别领域[50]。相比以上两种深层神经网络,CNN 可在保证识别率的同时,还能大大降低模型的复杂度,从而降低语音识别过程中对最开始的语音特征提取的依赖。值得注意的是,二维图像作为 CNN 的输入数据,两个维度上的特征物理意义一样,但将语音作为二维特征输入时,其物理意义不相同。文献[51]提到,将语音的二维特征分为时域和频域两个维度,此时 CNN 中的 C 层可看做是通过滤波器对局部频域特征的观察,进而抽取局部有用信息;而 S 层是在相邻两个 feature map 的输出节点中选择最大值作为输出;之后与图像一样,最终需通过一个全连接层得到各个状态的分类后验概率来得到分类结果。2012 年,多伦多大学初步建立了 CNN 用于语音识别的模型结构,并同传统训练算法的 DNN 相比取得相对 10% 的性能提升[26]。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 CNN 相对传统算法的 DNN 建模的有效性。CNN 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特征的鲁棒性。在文献[51]中,深入分析了 CNN 中卷积层和聚合层的不同结构对识别性能的影响情况,在标准英语连续语音识别库以及汉语电话自然口语对话数据集上,对卷积神经网络的输入特征、卷机器个数与尺寸、模型规模和计算量等做了详细的对比实验,并且还与深层神经网络进行了对比实验。实验中卷积神经网络的训练采用一层卷积层 + 子采样层 + 一层全网络层的结构,而为了保证与卷积神经网络的层数一致,训练 DNN 模型采用的是两个隐含层结构,结果表明 CNN 的识别率更高,泛化能力更强。而在文献[27]中,提出了一种新的设计 CNN 池化层的策略,在音素识别任务上比以前所有的 CNN 效果都要好。

在最新的研究[52]中,RNN 充当后端识别器,DNN 的高维输出不经过降温而直接将其作为特征输入给该识别器。研究表明,从 RNN 序列识别器的识别精度来看,使用 DNN 最高隐层作为特征相比其他隐层或输出层的效果更好。

DNN 强大的建模能力降低了系统的复杂性,截止 2013 年 ICASSP 会议[53],全世界至少有 15 个主要的语音识别团队的实验证明了在大规模语音识别任务上使用 DNN 的有效性,以及用原始语音序列的频谱特征(而不是 MFCC)可以得到更好的结果。这些团队包括了著名的工业界语音实验室,如 Microsoft、IBM、Google、讯飞和百度等。

2.2 利用深度学习网络进行声学建模

深度学习应用于声学建模很常见,因为为了解决语音识别中非常有挑战性的声学建模问题,出于经济层面的考虑,构建全世界所有语种的语音识别系统,瓶颈在于缺乏标注的语音数据,所以将深度学习神经网络用于声学建模当中。在文献[54]中使用 DNN 提高英语 ASR 系统的声学模型,错误率比传统 ASR 系统下降了 33%。早期的 DNN-HMM 构架[55]是在 NIPS 研讨会[56]上提出的,该架构由多伦多大学和微软研究院的语音研究者建立。基于 GMM-HMM 的声学模型是目前对 HMM 输出概率进行建模的主流方法,该方法主要是基于上下文相关的浅层、扁平的 GMM 和 HMM 生成式模型,但当面对更加复杂的语音识别环境时,GMM 逐渐显示出建模能力不足的问题,因为 GMM 本质上仍然属于浅层结构。

在文献[38]中,使用 5 层 DNN(在文献中称为 DBN)替换 GMM-HMM 系统中的混合高斯模型(GMM),并以单音素(monophone)状态作为建模单元。尽管单音素比三音素(triphone)的表征能力差一些,但使用单音素 DBN-HMM 构架的方法却比当时最先进的三音素 GMM-HMM 系统识别率更高,并且 DBN 对 HMM 中后验概率的估计不需要很苛刻的数据分布假设,条件更宽泛。在文献[45]中,研究了基于 DNN-HMM 的声学建模方法,它是在 Kaldi 开源语音识别平台上分别实现了基于 GMM-HMM 和 DNN-HMM 的声学建模,并且在 RM 语音库上通过实验证明了应用后者的识别系统比前者在词识别错误率上相对下降了 30%,而此文献中 DNN 各层节点的训练方式与 RBM 的训练方式类似。

与其他分类器相比,DNN 最主要的优势是加强了语音帧与帧之间的联系。在文献[57]中,进行了一个关于对 DNN 声学模型设计里哪一方面最重要的调查,用几个指标来比较不同的 DNN 分类器,分析影响性能的因素,从而观察这些因素对最终的语音识别的词错误率的影响。实验中发现,影响整个网络最重要的因素是层数。当 DNN 层数增加到某个值后再增加时,不仅对 DNN 分类器的性能没有提高,反而降低了,最终表明 3 ~ 5 个隐层的 DNN 结构大小是足够的。在文献[58]中,研究者用一个单独的 DNN 估计语音每一帧的语音序列后验概率,并用一个加权有限状态的传感器为解码器来估计分析语音,实验表明在不同的信噪比下,系统的最佳平均词错误率为 18.8%,比现在最先进的 IBM 系统还降低了 2.8%。

在文献[59]中提出了一种新的深度学习算法——深凸网络(deep convex network,DCN),用来解决语音识别中难以处理大量数据的可扩展性问题。在 DCN 的整个学习过程中,是基于成批式处理模式而不是随机的,在 MNIST 和 TIMIT 任务中的实验结果表明 DCN 更优于 DBN,这不仅表现在可扩展性训练和只有 CPU 的计算中,还表现在对这两个任务中语音数据的分类准确性上。还有在文献[5]中,CD-DNN-HMM 与传统的用序列鉴别准则(sequence discriminative criteria)训练的 GMM-HMM 系统相比,Switchboard 对话人物上减少了三分之一的错误率。

文献[60]提出了一种新的 DNN 模型,这个新的 DNN 使用支持向量机(support vector machine,SVM)在顶层进行分类,而传统的 DNN 在顶层分类使用的是多项式逻辑回归(softmax)。文中在最大边际标准中将这两种算法呈现在帧级和序列级中去学习 SVM 和 DNN 的参数,在帧水平的训练中,新模型表现出与携带 DNN 特征的多分类 SVM 有关;而在序列级的训练中,它表现出与携带 DNN 特征和 HMM 状态转移特征的结构化 SVM 有关。它的解码过程类似于 DNN-HMM 混合系统,但具有帧级的后验概率由 SVM 分数代替。文中定义这种新的 DNN 为深度神经支持向量机(deep neural support vector machine,DNSVM),并且验证了其在 TIMIT 上的连续语音识别的有效性,这个新模型比传统的 DNN 模型错误率降低了 8% 以上。

在语音领域中,最有意思的多任务学习应用当属多语种或交叉语种的语音识别,即不同语言的语音识别被当做不同的任务。据报道,2012 年 11 月,微软在天津演示了一个全自动同声翻译系统,其关键技术就是 DNN[61]。最近的几篇文献[53,62,63]中也提出了非常相近的、具有多任务学习能力、用于多种语音识别的深度神经网络构架。这一构架的思想是:通过适当的学习,深度神经网络中由低到高的隐层充当着复杂程度不断增加的特征变换,而这些变换共享跨语言声学数据中共有的隐藏因素;网络最后一个 softmax 层充当着一个对数线性(log-liner)分类器,利用了最顶端隐层所表示的最抽象的特征向量。尽管对数域的线性分类器对不同语言必要时可以分开,但特征转换仍然可以在跨语言之间共享[64]。它表明了可以从一个现有的多语种 DNN 中快速构建出一个性能良好的新语种 DNN 识别器。最大的好处莫过于只需要目标语言少量的训练数据,就可降低无监督预训练阶段的需求,并且可减少迭代次数。

而对于 CNN 的声学建模,在 2014 年,IBM 的沃森研究中心 Sainath[50]的工作结果显示 DNN 比以往过去的 GMM-HMM 模型有 8% ~ 15% 的提升,而 CNN 相比 DNN 有更强的适应能力,同时还具有数据平移不变性的特性。在文献[65]中最新的研究实验结果表明,将 CNN 应用到 LCVSR 中,比 DNN 的错误率降低了 1.8%。并且在文献[66,67]中对于远距离的语音识别,测试结果都显示 CNN 比 DNN 更有效。但在文献[68,69]的实验中,研究人员发现进行带噪语音识别时,CNN 的表

现并不理想。

而对于 RNN,文献[70]提到采用双向 RNN,对 2007 IBM GALE 的识别错误率仅为 12.6%。在文献[71]中,RNN 使用长期—短期的记忆(LSTM)结构展示了在语音识别上最先进的性能。它是将 RNN 中 softmax 层替换成 SVM,RNN 和 SVM 的参数学习是使用序列级的最大边际标准,而不是交叉熵。所以这种结果模型被定义为递归 SVM。

大量研究表明,将深度学习应用于提取语音特征和取代 HMM 中的 GMM 模型非常成功,并且将深度学习应用到语音识别当中,是将深度学习应用到工业领域的第一个成功案例。研究人员将不断研究新的深度学习神经网络模型取代整个语音识别系统来构建更好的语音识别系统。

3　结束语

本文简要介绍了深度学习发展历史以及目前应用最多的几种深度学习模型,并阐述了将这几种深度学习模型用于语音识别领域的发展与现状。深度学习的研究仍处于发展阶段,主要面临的问题有:a)训练通常需要解决一个高度非线性优化问题,导致在训练网络过程中容易陷入许多局部极小;b)如果训练太长,会导致结果容易过拟合化。而利用深度神经网络解决鲁棒性问题是现在语音识别领域中最热门的话题,在实际应用中的带噪语音识别率仅为 70% 左右,所以至今没有一个稳定、高效、普适的系统可以对带噪语音的识别率达到 90% 以上。对于未来语音识别的研究,最好的发展方向就是仿脑和类脑计算,只有不断符合人脑的语音识别的特性,才能将语音的识别率提高到令人满意的程度,然而现有的深度学习技术对于达到这一要求还远远不够。如何将深度学习更好地应用和达到市场对高效语音识别系统的需求将是一个值得继续重点关注的问题。

参考文献:

[1] 赵力. 语音信号处理[M]. 2 版. 北京:机械工业出版社,2011.

[2] 刘幺和,宋庭新. 语音识别与控制应用技术[M]. 北京:科学出版社,2008.

[3] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006,18(3):1527-1554.

[4] Bengo Y, Courcille A, Vincent P. Representation learning: a review and new perspectives[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2013,35(8):1798-1828.

[5] Dahl G, Yu Dong, Deng Li, *et al*. Context-dependent pretrained deep neural networks for large vocabulary speech recognition[J]. IEEE Trans on Audio, Speech, and Language Processing, 2012,20(1):30-42.

[6] Hinton G E, Deng Li, Yu Dong, *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.

[7] 余凯,贾磊,陈雨强,等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展,2013,50(9):1799-1804.

[8] 刘建伟,刘媛,罗雄麟. 深度学习研究进展[J]. 计算机应用研究,2014,31(7):1921-1930.

[9] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009,2(1):1-127.

[10] 戴武昌, 王建国, 徐天锡. 基于神经网络的蓄电池荷电状态估算[J]. 东北电力大学学报, 2016,36(5): 2-3.

[11] Bengio Y,Lamblin P,Popovici D,*et al*. Greedy layer-wise training of deep networks[C]//Proc of the 19th International Conference on Neural Information Processing Systems. Cambridge:MIT Press, 2007: 153-160.

[12] Schölkopf B, Platt J, Hofmann T. Efficient learning of sparse representations with an energy-based model[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2006: 1137-1144.

[13] Hinton G E, Salakhutdinov R. Reducing the dimensionality of data with neural networks[J]. Science, 2006,313(5786): 504-507.

[14] Bengio Y, Dwlalleau O. On the expressive power of deep architectures [C] //Proc of the 22nd International Conference on Algorithmic Learning Theory. 2011: 18-36.

[15] 郭丽丽, 丁世飞. 深度学习研究进展[J]. 计算机科学, 2015,42(3):28-33.

[16] 刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. 计算机应用研究, 2014,31(7):1921-1928.

[17] Hinton G E. A practical guide to training restricted Boltzmann machines[J]. Momentum, 2010,9(1): 599-619.

[18] 张建明, 詹智财, 成科扬,等. 深度学习的研究与发展[J]. 江苏大学学报: 自然科学版, 2015,36(2): 191-200.

[19] Cho K Y. Improved learning algorithms for restricted Boltzmann machines[D]. Espoo: Aalto University, 2011.

[20] 梁静. 基于深度学习的语音识别研究[D]. 北京: 北京邮电大学, 2014.

[21] Alpayd E. 机器学习导论[M]. 范明, 等译. 北京: 机械工业出版社, 2009.

[22] Larochelle H, Bengio Y, Louradour J, *et al*. Exploring strategies for training deep neural networks [J]. Journal of Machine Learning Research, 2009,10(12): 1-40.

[23] Vincent P, Larochelle H, Lajoie I, *et al*. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010,11(6): 3371-3408.

[24] 刘进峰. 一种简洁高效的加速卷积神经网络的方法[J]. 科学技术与工程, 2014, 14(33): 240-244.

[25] Abdel-Hamid O, Deng Li, Yu Dong. Exploring convolutional neural network structures and optimization techniques for speech recognition [J]. Interspeech,2013,58(4): 1173-1175.

[26] Abdel-Hamid O, Mohamed A, Jiang Hui, *et al*. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition [C]//Proc of International Conference on Acoustics, Speech, and Signal Processing. 2012: 4277-4280.

[27] Deng Li, Abdel-Hamid O, Yu Dong. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion [C]// Proc of International Conference on Acoustics, Speech, and Signal Processing. 2013: 6669-6673.

[28] 段艳杰, 吕宜生, 张杰, 等 . 深度学习在控制领域的研究现状与展望[J]. 自动化学报, 2016,42(5): 644-645.

[29] Bengio Y. Deep learning of representations: looking forward[M]// Statistical Language and Speech Processing. Berlin: Springer, 2013: 1-37.

[30] Sutskever I. Training recurrent neural networks[D]. Toronto: University of Toronto, 2013.

[31] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 北京: 清华大学出版社, 2004.

[32] Yu Dong,Deng Li. 解析深度学习——语音识别实践[M]. 俞凯, 钱彦旻,等译. 北京:电子工业出版社,2016.

[33] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex[J]. Journal of the Optical Society of America a Optics Image Science & Vision, 2003,20(7): 1434-1448.

[34] Deng Li. Industrial technology advances: deep learning from speech recognition to language and multimodal processing [J]. APSIPA Trans on Signal and Information Processing, 2016(5).

[35] Mohamed A, Yu Dong, Deng Li. Investigation of full-sequence training of deep belief networks for speech recognition[C]//Proc of Conference of the International Speech Communication Association. 2010: 2846-2849.

[36] Deng Li, Yu Dong. Deep learning for signal and information processing[R]. [S. l.]: Microsoft Research, 2013.

[37] Dahl G E, Yu Dong, Deng Li, *et al*. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2011: 4688-4691.

[38] Mohamed A, Dahl G, Hinton G. Acoustic modeling using deep belief networks[J]. IEEE Trans on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

[39] Sivaram G S V, Hermansky H. Sparse multi-layer perceptron for phoneme recognition[J]. IEEE Trans on Audio, Speech, and Language Processing, 2012, 20(1): 23-29.

[40] Jaitly N, Hinton G. Learning a better representation of speech soundwaves using restricted Boltzmann machines[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2011: 5884-5887.

[41] Seide F, Li Gang, Chen Xie, *et al*. Feature engineering in context-dependent deep neural networks for conversational speech transcription[C]//Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. 2011: 24-29.

[42] Deng Li. Switching dynamic system models for speech articulation and acoustics[M]// Mathematical Foundations of Speech and Language Processing. New York: Springer, 2004: 115-133.

[43] Jaitly N, Nguyen P, Vanhouche V. Application of pretrained deep neural networks to large covabulary speech recognition[C]// Proc of Interspeech. 2012.

[44] Mohamed A R, Dahl G E, Hinton G E. Deep belief networks for phone recognition[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2011: 5060-5063.

[45] 李晋辉, 杨俊安, 王一. 一种新的基于瓶颈深度信念网络的特征提取方法及其在语种识别中的应用[J]. 计算机科学, 2014, 41(3): 263-266.

[46] You Yongbin, Qian Yanmin, He Tianxing, *et al*. An investigation on DNN-derived bottleneck features for GMM-HMM based robust speech recognition[C]//Proc of IEEE China Summit and International Conference on Signal and Information Processing. [S. l.]: IEEE Press, 2015.

[47] Qian Yanmin, He Tianxing, Deng Wei, *et al*. Automatic model redundancy reduction for fast back-propagation for deep neural networks in speech recognition[C]//Proc of International Joint Conference on Neural Networks. [S. l.]: IEEE Press, 2015.

[48] Liu Yuan, Fu Tianfan, Fan Yuchen, *et al*. Speaker verification with deep features[C]//Proc of International Joint Conference on Neural Networks. 2014: 747-753.

[49] Imseng D, Motlicek P, Garner P, *et al*. Impact of deep MLP architecture on different modeling techniques for under-resourced speech recognition[C]//Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. 2013: 332-337.

[50] Sainath T N. Improvements to deep neural networks for large vocabulary continuous speech recognition tasks[R]. [S. l.]: IBM Thomas J. Watson Research Center, 2014.

[51] 张晴晴, 刘勇, 潘接林, 等. 卷积神经网络在语音识别中的应用[J]. 工程科学学报, 2015, 37(9): 1217-1217.

[52] Chen Jianshu, Deng Li. A primal-dual method for training recurrent neural networks constrained by the echo-state property[C]//Proc of International Conference on Learning Representations. 2013.

[53] Heigold G, Vanhoucke V, Senior A, *et al*. Multilingual acoustic models using distributed deep neural networks[C]//Proc of International Conference on Acoustics Speech and Signal Processing. 2013: 8619-8623.

[54] Nguyen Q B, Vu T T, Chi M L. Improving acoustic model for English ASR System using deep neural network[C]// Proc of IEEE RIVF International Conference on Computing & Communication Technologies: Research, Innovation, and Vision for the Future. 2015.

[55] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]//Proc of NIPS Workshop. 2010.

[56] Deng Li, Yu Dong , Hinton G . Deep learning for speech recognition and related applications[C]//Proc of NIPS Workshop. 2009.

[57] Maas A L, Qi Peng, Xie Ziang, *et al*. Building DNN acoustic models for large vocabulary speech recognition[J]. Computer Speech & Language, 2015, 41(1): 195-213.

[58] Weng Chao, Yu Dong, Seltzer M L, *et al*. Deep neural networks for single-channel multi-talker speech recognition [J]. IEEE/ACM Trans on Audio Speech & Language Processing, 2015, 23(10): 1670-1679.

[59] Yu Dong, Deng Li. Deep convex net: a scalable architecture for speech pattern classification[C]//Proc of the 12th Annual Conference of International Speech Communication Association. 2011: 2285-2288.

[60] Zhang Shixiong, Liu Chaojun, Yao Kaisheng, *et al*. Deep neural support vector machines for speech recognition[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015: 5885-5889.

[61] Markoff J. Scientists see promise in deep-learning programs[N]. The New York Times, 2012-11-23.

[62] Deng Li, Li Jinyu, Huang J T, *et al*. Recent advances in deep learning for speech research at Microsoft[C]//Proc of International Conference on Acoustics Speech and Signal Processing. 2015: 8604-8608.

[63] Huang J T, Li Jinyu, Yu Dong, *et al*. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers[C]// Proc of IEEE International Conference on Acoustics, Speech & Signal Processing. 2013: 7304-7308.

[64] Deng Li, Yu Dong. Deep learning: methods and applications[M]. [S. l.]: Microsoft Research, 2016.

[65] Zhang Qingqing, Liu Yong, Wang Zhichao, *et al*. The application of convolutional neural network in speech recognition[J]. Journal of Network New Media, 2014, 22(10): 1533-1545.

[66] Sainath T N, Mohamed A R, Kingsbury B, *et al*. Deep convolutional neural networks for LVCSR[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 8614-8618.

[67] Huang J T, Li Jinyu, Gong Yifan. An analysis of convolutional neural networks for speech recognition[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015.

[68] Palaz D, Magimai-Doss M, Collobert R. Convolutional neural networks-based continuous speech recognition using raw speech signal[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015: 177-181.

[69] Chan W, Lane I. Deep convolutional neural networks for acoustic modeling in low resource languages[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015: 2056-2060.

[70] Arisoy E, Sethy A, Ramabhadran B, *et al*. Bidirectional recurrent neural network language models for automatic speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2015: 5421-5425.

[71] Zhang Shixiong, Zhao Rui, Liu Chaojun, *et al*. Recurrent support vector machines for speech recognition[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2016.