卷积神经网络在语音识别中的应用*

张晴晴 刘 勇 王智超 潘接林 颜永红

(中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)

摘要: 研究了使用卷积神经网络构造模式分类器,并用于连续语音识别的研究。CNNs 相比于广泛使用于语音识别中的深层神经网络(Deep Neural Network,DNNs),能在保证性能的同时,大大压缩模型的尺寸。在标准语音识别库 TIMIT 上的实验结果证明,相比传统 DNN 模型,CNN 模型的识别性能更好,同时其模型规模和计算量都有明显降低。

关键词: 卷积神经网络,连续语音识别,权值共享

The Application of Convolutional Neural Network in Speech Recognition

ZHANG Qingqing, LIU Yong, WANG Zhichao, PAN Jielin, YAN Yonghong

(Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: Convolutional Neural Networks (CNNs) are investigated for continuous speech recognitions in the paper. Compared to Deep Neural Networks (DNNs), which have been proven to be successful in many speech recognition tasks nowadays, CNNs can reduce the NN model sizes significantly, and at the same time achieve even better recognition accuracies. Experiments on standard speech corpus TIMIT showed that CNNs outperformed DNNs in accuracy.

Keywords: Convolutional Neural Networks, Continuous speech recognition, Weight – sharing

1 引言

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 DNN 在语音识别领域复兴的热潮。目前包括微软、IBM、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 DNN 的研究[2]。

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\\图像分割和文本检测中,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 CNNs(Convolutional Neural Networks,CNNs)[3]。

本文于 2014 – 07 – 10 收到。

* 基金项目:国家自然科学基金(编号:11161140319,91120001,61271426),中国科学院战略性先导科技专项(面向感知中国的新一代信息技术研究,编号:XDA06030100,XDA06030500),国家 863 计划(编号:2012AA012503),和 中科院重点部署项目(编号:KGZD – EW – 103 – 2)基金资助。

CNNs 的权值共享网络结构使之更类似于生物神经网络，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，所以在图像处理中得到了广泛的使用。在本研究中，CNNs 被引入连续语音识别中，并和目前广泛使用的 DNNs 模型进行了对比。相比 DNNs，CNNs 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，CNNs 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 TIMIT① 上面进行了实验，对 CNNs 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介绍，实验结果证明相比传统 DNN，CNN 结构在获得更好识别性能的同时，降低了模型规模和计算量。

2 卷积神经网络

CNNs 由一组或多组卷积层 convolutional layer + 采样层 pooling layer 构成[3]。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。采样层通过对卷积层的输出结点做固定窗长的采样，减少下一层的输入结点数，从而控制模型的复杂度。一般采样层采用最大采样算法（max pooling），即对固定窗长内的结点选取最大值进行输出。最后，通过全网络层将采样层输出值综合起来，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[4]。CNN 相比 DNN 等神经网络结构，引入了三个重要的概念：局部卷积、采样、权值共享。在本章中，我们将详细介绍语音识别中使用的 CNNs 的结构，以及在语音识别中引入 CNNs 时对这三个重要概念的理解。

图 1 给出了 CNN 网络用于语音识别声学建模时，典型的卷积层和采样层的结构。当二维图像作为 CNN 的输入时，两个维度上特征的物理意义是完全一样的。而将语音看做二维特征输入时，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 DNN 上实验证明，多帧串联的长时特征对模型性能的提高很重要，在 CNN 的输入特征上，我们也保留了该方法，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 CNN 的第一维特征。在 CNN 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）。CNN 中卷积层的物理意义是：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）。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，通过式（1）计算得到卷积器的输出：

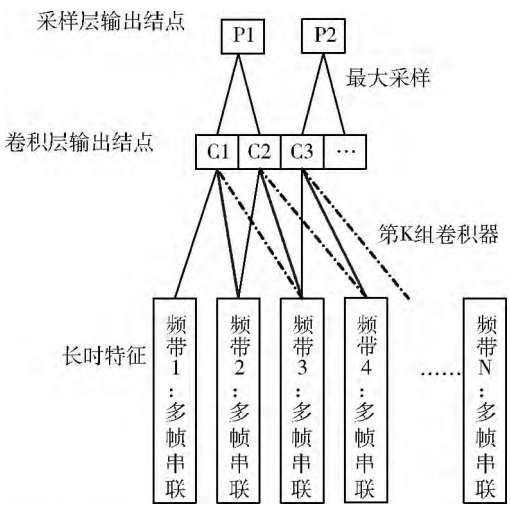

图 1 CNNs 中卷积层和最大采样层的示例图

$$C_{i,k} = \theta\left(\sum_{b=1}^{s-1} w_{b,k} v_{b+i}^{T} + a_k\right) \quad (1)$$

式（1）中 k 表示第 k 个卷积器，v_{b+i}^{T} 表示第 i 组输入特征矢量，$w_{b,k}$ 为第 k 个卷积器的权值参数，s 则为卷积器的宽度，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 θ 得到卷积层的一个输出结点值，θ 一般选择反正切函数或 sigmoid 函数。

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

① http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

器的自由参数量之和,相比 DNN 等全网络连接结构,大大减少了模型的存储规模。

在卷积层之后,紧跟着的是采样层。在语音识别中,我们采用了最大采样算法(采样) 。以图 1 为例,我们从 C1 和 C2 这两个卷积层输出结点中选择最大值作为采样层的输出 P1。这样做的优点: 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,CNN 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,我们通过一个全网络层,将采样层的各个输出综合起来,最后通过输出层得到各个状态的后验概率。

3　实验结果和分析

3.1　实验条件

本文的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分类帧正确率(frame correct rate) 和最终的音素识别正确率(phone correct rate) 。我们使用了 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。各个集子间无说话人重叠。在特征提取部分,我们使用传统的 25ms 帧长 10ms 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 CNN 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。

CNN 的训练采用一层(卷积层 + 采样层) 和一层全网络层的结构。为了与之对比,训练了 DNN 模型,采用的是两个隐含层结构,保证和 CNN 的网络层数一致。CNN 和 DNN 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。在我们的实验中,为了直接观察 CNN 在声学建模上的性能,采用了不带语言模型的音素解码。

3.2　实验结果

表 1　TIMIT 测试集上 CNN 和 DNN 的参数和性能对比

		输入维数	卷积器个数	卷积器参数	Max pooling	全连接网络隐层	验证集帧正确率	测试集音素正确率
CNN	40 维	11 ×40	100 个	11 ×8	1 ×3	1024	47.6%	61.7%
	120 维	33 ×40	100 个	33 ×8	1 ×3	1024	53.6%	66.3%
DNN	40 维	11 ×40		1024		1024	46.3%	60.1%
	120 维	33 ×40		1024		1024	51.8%	64.6%

表 1 给出了 CNN 和 DNN 在不同条件下的性能对比结果。在特征方面,我们尝试了不使用\\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”。CNN 的结构为两个隐层,第一个隐层为卷积层 + 采样层,卷积器种类为 100 种,对应两种不同特征时的卷积器参数分别为 11 ×8 和 33 ×8 (见第二章说明) ,采样层为三个结点选择一个最大输出的方式。然后紧接一个 1024 结点的全网络隐层。基于 120 维特征输入的结构,CNN 的总模型大小为 2.6M,总计算量为 1.6M 次(矩阵乘法) 。DNN 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,DNN 的总模型大小为 10M,总计算量为 2.6M 次(矩阵乘法) 。对比看出,无论是模型规模还是实际计算量,CNN 都比 DNN 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,CNN 的帧正确率和音素正确率都稳定优于 DNN。并且,使用一阶和二阶差分特征会进一步提高模型性能。

4　结束语

本研究工作将卷积神经网络引入连续语音识别,并和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。在标准库上实验证明，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。

参 考 文 献

[1] Dahl, G. E., Dong Yu, Li Deng, et al.. Context – dependent pre – trained deep neural networks for large – vocabulary speech recognition [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1): 30 – 42

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

[3] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time – series [M], in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, Ed. MIT Press, 1995.

[4] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting [C], in Proceedings of CVPR'04. IEEE Press, 2004.

作者简介

张晴晴，女，1983 年 10 月出生，学历博士，副研究员，研究方向为语音识别声学建模和语言建模。

更正: 2014 年第 3 卷第 4 期中题目《地铁声学防撞辅助系统关键技术研制》的论文作者薛燕，闵静辉，李晓东，冯海泓，有误。作者应为薛燕，胡晓城，闵静辉，李晓东，冯海泓，特此更正。