

基于一维卷积神经网络的车载语音识别研究

朱锡祥,刘凤山,张超,吕钊,吴小培

(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要: 针对卷积神经网络(convolution neural networks,CNNs)是二维结构,不能很好地反映出语音信号的一维特性,因此,提出使用一维模型进行语音识别研究.其通过卷积核在时间轴上的移动,在保留频带相关性的同时可以更好的满足语音信号的时变性,进而提高识别性能.最后进行了车载语音识别对比实验,结果表明一维卷积神经网络的识别率较二维卷积神经网络提高了约 10%~20%,在噪声环境下的泛化性能也明显优于后者.

关键词: 卷积神经网络;语音识别;网络维度;卷积核;泛化性

中图分类号: TN912.34

文献标识码: A

文章编号: 1000-7180(2017)11-0021-05

Research on in-car Speech Recognition Based on One Dimensional Convolutional Neural Networks

ZHU Xi-xiang,LIU Feng-shan,ZHANG Chao,LV Zhao,WU Xiao-pei

(The Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: Convolution neural networks(CNNs) has been the architecture of traditional convolution neural networks is two-dimensional(2D), which can not reflect the one-dimensional characteristic of speech signal. Therefore, a one-dimensional(1D)architecture for speech recognition was proposed, which can better satisfy the temporal variation while retaining band correlation by convolution along the time axis. Experiments of in-car speech recognition demonstrate that 1D CNNs can significantly outperform the 2D CNNs, with recognition rate improvement of 10% to 20%, and the generalization performance in noisy environment of 1D CNNs is also significantly better than the latter.

Key words: convolution neural networks; speech recognition; dimension of network; convolution kernel; generalization

1 引言

语音识别作为人-机交互中的关键技术,一直受到广泛的关注和研究.伴随着信号处理和计算能力的不断进步,语音识别算法在近几十年里得到了迅速的发展.传统的动态时间规整(Dynamic Time Warping,DTW)、线性预测(Linear Prediction,LP)、隐马尔科夫模型(Hidden Markov Model,HMM)等识别算法在不同的时期和场景内都得到了一定程度的应用.人工神经网络(Artificial Neu-

ral Networks,ANNs)是一种模拟人脑神经元工作模式的计算模型,其在上世纪 80 年代提出的反向传播(Back Propagation,BP)算法^[1]的推动下得到了迅速的发展,并成功运用于语音识别领域^[2].卷积神经网络是一种前馈型神经网络,受到猫的视觉系统神经元局部敏感和方向选择特性的启发^[3],加入了局部感知和池化等能力,使得卷积神经网络对平移、倾斜、缩放等变形具有高度不变性,因此在 2012 年,卷积神经网络被应用于语音识别,并初步确定了网络的基本结构,实验证明其识别性能在深度神经网络

收稿日期:2017-03-01;修回日期:2017-03-28

基金项目:国家自然科学基金(61271352,61401002);安徽省自然科学基金(1408085QF125);安徽省科技攻关强警项目(1501b04220)

络的基础上得到了大幅提升^[4]. 此后 Sainath 等人对网络结构进行了调整和优化, 进一步提升了其识别性能^[5].

但是卷积神经网络在语音识别任务中一直使用的是传统的二维模型, 并不能很好的适应语音的一维特性. 为了解决这些问题, 本文提出了一维卷积神经网络模型, 并应用于车载语音识别任务中.

2 卷积神经网络

卷积神经网络除了输入输出层以外, 主要包括卷积层和采样(聚合)层:

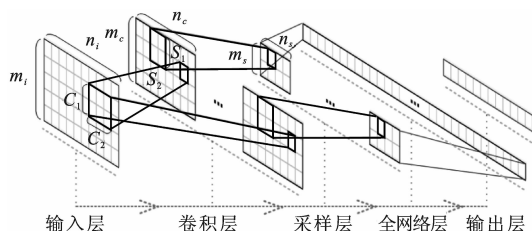


图1 卷积神经网络基本结构

图1中, m_i 、 n_i 和 m_c 、 n_c 分别为输入层和卷积层中特征图的维度大小, C_1 、 C_2 为卷积核的维度大小, S_1 、 S_2 是两个维度上降采样的比例. 从图中可以看出, 卷积层的特征图的大小由上一层特征图和卷积核的大小决定, 而采样层的特征图大小由上一层特征图和降采样的比例决定. 具体关系如下:

$$\begin{cases} m_c = m_i - C_1 + 1 \\ n_c = n_i - C_2 + 1 \\ m_s = m_c / S_1 \\ n_s = n_c / S_2 \end{cases} \quad (1)$$

卷积层通过卷积核对输入信号的卷积操作进行特征提取. 卷积核的形状和个数直接决定了网络的整体性能: 卷积核越小, 提取的特征会越细致, 但是会丧失输入信号的相关性信息, 使得网络的泛化性降低. 相反, 卷积核越大, 虽然会保留相关性信息, 但是会造成信号的细节缺失. 同时, 卷积核的形状也应与输入信号的特征相匹配, 例如在进行语音识别任务时, 为了更多地提取频带上的特征, 卷积核的一个维度通常与长时信号的帧数相对应^[4]. 卷积核个数的不同代表着输入信号的特征提取方式的不同, 即观察输入信号的角度不同. 一般情况下, 卷积核个数和网络整体性能呈正相关关系, 但是在增长到一定的上限后, 在增加网络计算规模的同时, 其给性能带来的提升会变得非常有限^[6]. 所以, 卷积层的设计是针对不同目标任务的调优过程.

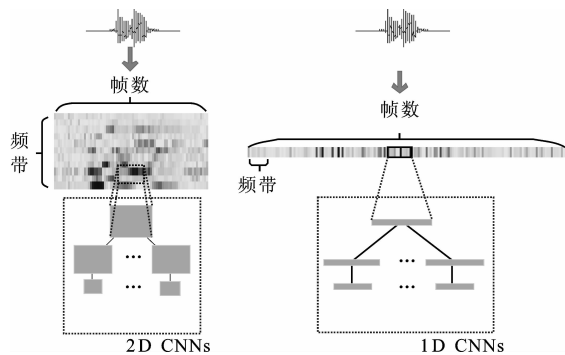
采样层会对卷积层提取出的特征图进行降采样, 采样方法有均值采样、最大值采样等, 在语音识别任务中一般采用最大值采样(max pooling). 采样比例也不宜过大, 虽然采样比例的增大会提高网络泛化能力, 降低网络规模, 但是过大的降采样比例会丢失大量特征信息, 造成网络性能的下降, 所以采样比例的选择要适中. 这种对提取出的特征进行高层聚合的行为是除了局部卷积以外, 提高网络泛化能力的第二个重要途径.

一般在输出层之前会添加一个全网络层, 用以将之前的局部观察结果进行整合连接, 进而在输出层得出分类结果或后验概率.

3 一维卷积神经网络

3.1 一维网络结构

卷积神经网络起初被广泛应用于图像等领域, 网络的输入大多是图片等二维矩阵, 如图1所示, 卷积核、特征图等网络内部结构也都是二维的. 随后卷积神经网络被应用于语音识别任务, 虽然语音是典型的一维信号, 但是为了适应二维网络结构以及构建频带上的局部观察能力, 传统的做法是将语音分帧、提取特征参数后并排排列, 构成长时特征, 此时的语音信号虽然是二维的, 但是两个维度的含义并不相同, 一个是时间特征, 即帧数, 另一个是特征参数, 如图2(a)所示.



(a) 二维卷积神经网络结构 (b) 一维卷积神经网络结构

图2 语音识别任务中的两种卷积神经网络结构

在保留卷积神经网络的局部观察、权值共享以及高层聚合等特性^[3]的同时, 为了适应语音信号的一维特性, 本文构建出一维卷积神经网络, 用于语音信号的处理, 如图2(b)所示.

一维卷积神经网络中的输入是一维向量, 所以网络内部的卷积核、特征图也都是二维的, 即图1中的 m_i 、 m_c 、 m_s 和 C_1 、 S_1 均等于1. V_{ij}^y 表示第 i 层第 j 个特征图上第 y 个位置的值, 其可通过上一层的一

维向量和一维卷积核进行卷积得到:

$$V_{ij}^y = \text{sigmoid}(b_{ij} + \sum_m \sum_{l=0}^{L_i} \omega_{ijm}^l V_{(i-1)m}^{y+l}) \quad (2)$$

式中, $\text{sigmoid}(\cdot)$ 为传递函数,也可选用 $\tanh$ 、 Relu 等其他传递函数. b_{ij} 是此特征图的偏置, m 为在 $(i-1)$ 层中连接到此特征图的特征图集合的序号, ω_{ijm}^l 表示连接到序号为 m 的特征图的卷积核中第 l 位置的值, L_i 代表第 i 层中卷积核的长度.

采样层和二维卷积神经网络类似,仅需要在上一层的一维特征图上进行均值抽样或者最大值抽样.而全网络层对上一层进行拼接时不再需要将特征图重置成一维向量,因为在一维网络结构中,所有特征图都是一维结构,直接前后连接即可.

3.2 输入和分析

语音识别任务中卷积神经网络的传统结构是二维的,网络的输入、卷积核、特征图都是二维矩阵.此时网络的输入为不同帧的特征参数并列放置产生.假设提取的特征参数为 12 维的 MFCC,帧数为 11,则二维卷积神经网络的输入为 12×11 的特征参数矩阵.在这种情况下,卷积核在特征图上的卷积将重点发生在频带,即提取的是频带上的局部特征分布^[7].虽然这对分布在不同频带上的噪声有一定的鲁棒性^[8],但是并不能凸显出一维语音信号的时变特性,以及频带上的相关性.

针对语音信号的一维特性和时变性,本文将输入置为一维向量.具体的做法是将语音信号提取出的特征参数按照帧顺序,首尾相连.例如将前文提到的 11 帧的 12 维 MFCC 首尾相连,构成 1×132 的一维向量作为输入.

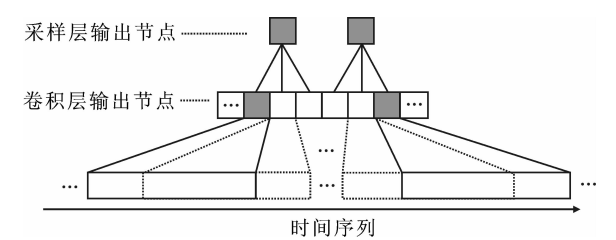


图3 一维卷积神经网络的短时特性

如图 3 所示,此时的卷积核相当于在时间轴上滑动的观察窗,卷积操作提取的是时间轴上的短时特征,这些短时特征在采样层进一步聚合,选取其中一部分保留.通过这两步操作,可以很好地分析和保留住时间序列上的局部特征,满足语音信号的时变性.

在二维网络中,卷积的部分主要发生在输入的

特征矩阵的频带维度,虽然在部分识别任务中会有后续的隐马尔可夫模型保证时间序列的相关性,但是对于简单的识别场景,这样的网络结构过于复杂.所以本文选择一维模型进行语音识别任务,在降低识别模型的复杂度的同时,其通过卷积核在时间轴上的观察能力,可以在保留频带相关性和完整性的同时,提取信号的短时特征,这将有利的提高网络的识别性能.

4 实验结果和分析

为了讨论一维卷积神经网络和二维卷积神经网络的识别性能.从以下四个方面分别作出对比实验.首先会展示不同时间规整帧数对实验结果的影响,在对比两种网络的同时,消除前端处理部分的变化对后续实验的影响.随后通过网络的收敛性对比,对两种网络的识别性能进行初步讨论,最终通过卷积核形状和泛化性的对比实验,完成两种网络模型在不同环境下的识别性能的综合对比.以下实验结果中的识别率均为统计平均结果.

4.1 时间规整的影响

语音样本在输入到卷积神经网络之前,需要把特征参数规整到统一大小,即将拥有不同帧数的样本特征全部规整到同样的帧数.下面对比了不同时间规整帧数对识别性能的影响:

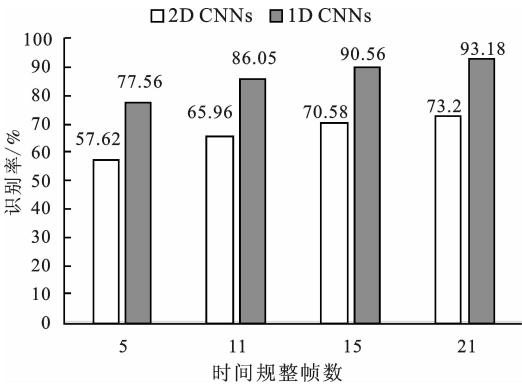


图4 时间规整对识别率的影响

通过图 4 可以看出,随着时间规整帧数的增加,网络的输入携带的信号特征越来越丰富,网络的识别率因此得到提升.同时,对比二维卷积神经网络和一维卷积神经网络可以看出,无论时间规整帧数如何变化,一维卷积神经网络的识别率均要明显高于二维卷积神经网络.

4.2 收敛性比较

神经网络的训练实际上是通过输出与标签之间的误差对网络参数进行调整以进一步降低误差的调

优过程. 因此, 误差的收敛性直接代表了神经网络的整体性能是否优良. 下面通过一维卷积神经网络与二维卷积神经网络的收敛曲线对二者的网络性能进行比较.

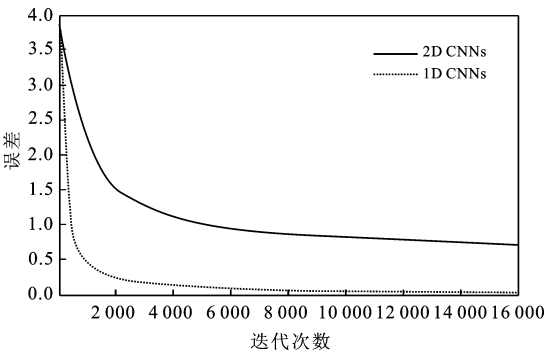


图 5 误差收敛趋势

通过图 5 中的收敛曲线, 可以很明显地看出一维卷积神经网络的误差要比二维神经网络收敛地更迅速, 而且从最终的收敛结果来看, 一维卷积神经网络的误差也要明显小于二维卷积神经网络. 因此, 通过网络误差的收敛性判断, 一维卷积神经网络要明

显优于二维卷积神经网络.

本次实验中采用的卷积核大小相同, 均为时间规整后的帧数 21, 但是在两种网络中是有区别的, 主要在于: 二维卷积神经网络中卷积核在频带上滑动, 而一维卷积神经网络中的卷积核在时间轴上滑动. 关于卷积核的区别及其在两种网络中的影响将在下一小节中重点讨论.

4.3 卷积核

为了比较不同卷积核形状对两种卷积神经网络的影响, 卷积核的大小在两种网络中保持一致, 仅仅将卷积核的形状改变为对应的一维和二维形式. 如表 1 中编号为 3 的卷积核参数, 1×63 代表着卷积核的长度为 63, 即其在时间轴上的观察窗口的长度为 63, 而对应的 3×21 的卷积核代表着卷积核在频带上观察的窗口长度为 3, 在时间轴上串联了 21 帧 (传统的语音识别任务中, 卷积核细节观察的是频带上的信息, 所以一般将卷积核的一个维度置为帧数^[6], 这样可以保证提取的是频带特征).

表 1 卷积核参数设置

参数编号	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
卷积核大小	21	42	63	84	105	126	147
卷积核形状	1D CNNs	1×21	1×42	1×63	1×84	1×105	1×147
	2D CNNs	1×21	2×21	3×21	4×21	5×21	7×21

从图 6 表示的实验结果中可以看出, 二维卷积神经网络的卷积核在频带上的观察窗口从 1 增加到 2, 增加了频带之间的相关性信息, 使得对应的识别率提高了近 7%, 但是随后识别率逐步下降, 这是因为随着观察的窗口的增大, 会逐步丢失细节信息. 在一维卷积神经网络中, 观察窗最小的时候, 就已经包含了频带的相关性信息, 识别率达到最高, 其后随着观察窗的增大, 识别率略有下降, 但并没有明显的波动. 在卷积核形状的改变过程中, 一维卷积神经网络

的识别率始终保持着比二维卷积神经网络高出约 10%~20%.

4.4 泛化性

泛化性是神经网络在实际应用中的重要性能指标之一, 一般指网络对训练集以外的测试数据做出正确反应的能力. 下面将分别对比一维卷积神经网络和二维卷积神经网络在不同测试集下的泛化能力.

表 2 中统计了两种卷积神经网络使用不同形状卷积核在三种测试集上的识别率, 其中带噪混合测试集是通过混合测试集卷积时速 80 km/h 的车载噪声仿真产生的, 其信噪比约为 -5.9 db. 卷积公式如下:

$$Y = h_1 * S + h_2 * N \tag{3}$$

Y 为加噪后语音, S 为纯净语音, N 为噪声, h_1 、 h_2 代表混合滤波器系数:

$$h_1 = [0.10 \quad 0.55 \quad -0.43 \quad 0.73 \quad 0.26 \quad -0.38]$$
$$h_2 = [0.43 \quad -0.26 \quad 0.88 \quad 0.03 \quad 0.63 \quad 0.46]$$

(4)

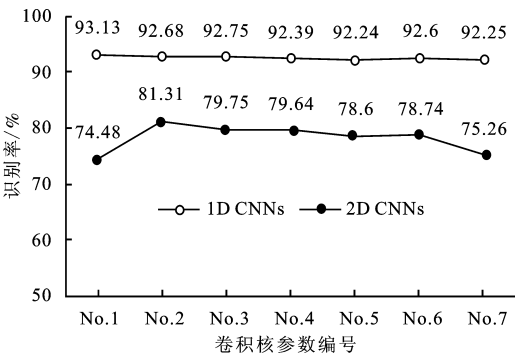


图 6 卷积核形状对识别率的影响

表 2 两种卷积神经网络的泛化性对比

网络维数	测试集	识别率/%					
卷积核参数编号		No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
1D	相同说话人	95.38	95.10	95.48	94.88	94.28	93.55
	不同说话人	90.6	90.45	90.93	90.58	91.03	90.48
	带噪混合	39.5	50.45	50.01	52.85	57.01	58.29
2D	相同说话人	74.75	79.45	80.03	79.88	79.60	79.38
	不同说话人	74.78	81.03	80.03	79.85	78.38	78.03
	带噪混合	29.03	47.45	41.51	46.36	52.59	54.28

从表中的统计结果可以看出,虽然两种卷积神经网络在面对不同的测试集时,识别性能都有不同程度的下降,尤其是在测试集带噪的情况下,识别率下降了约 40%,但是一维卷积神经网络的识别性能始终高于二维卷积神经网络. 在相同说话人和不同说话人这两种不含噪的纯净测试集中,一维卷积神经网络的识别率平均要比二维卷积神经网络高出约 14%,而在带噪测试集中,一维卷积神经网络的识别率平均要比二维卷积神经网络高出约 6%,识别率差距的缩小是因为车载噪声是一种低频噪声,其在频带上的分布是局部的,这非常有利于二维卷积神经网络进行频带上的局部特征提取. 相应的,在低噪环境中,语音的时变性和频带的相关性的重要性会更加凸显,此时一维卷积神经网络会更具优势.

5 结束语

本文针对语音信号的一维特性,提出使用一维模型进行语音识别研究,其卷积核作为观察窗,在语音信号的时间轴上滑动,用以提取信号的局部特征,保证了语音信号的时变性和频带的相关性、完整性. 在孤立词识别实验中,分别对比了一维卷积神经网络和传统的二维卷积神经网络在不同规整帧数、不同卷积核形状下的识别率,以及两者在不同测试集下的泛化性,结果表明一维卷积神经网络在以上不同环境中的识别性能均要明显优于二维卷积神经网络. 关于不同噪声强度对两种网络模型的具体影响仍然值得继续讨论和研究.

参考文献:

[1] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323 (6088): 533-536.

[2] Dahl G E, Yu D, Deng L, 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.

[3] Lecun Y, Bengio Y. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.

[4] Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]// Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. Japan, Kyoto, IEEE, 2012: 4277-4280.

[5] Sainath T N, Kingsbury B, Saon G, et al. Deep convolutional neural networks for large-scale speech tasks[J]. Neural Networks, 2015(64): 39-48. doi: 10.1016/j.neunet.2014.08.005.

[6] 张晴晴,刘勇,潘接林,等. 基于卷积神经网络的连续语音识别[J]. 工程科学学报, 2015, 37(9): 1212-1217.

[7] Abdel-Hamid O, Mohamed A, Jiang H, et al. Convolutional neural networks for speech recognition[J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2014, 22(10): 1533-1545.

[8] Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition [J]. Interspeech, 2013, 58(4): 1173-1178.

作者简介:

朱锡祥 男,(1991-),硕士研究生. 研究方向为语音识别、神经网络等. E-mail:hixxZhu@126.com.

刘凤山 男,(1993-),硕士研究生. 研究方向为语音增强、信号处理等.

张超 男,(1983-),博士,讲师. 研究方向为智能信息处理、视频/图像处理等.

吕钊 男,(1979-),博士,副教授、硕士生导师. 研究方向为智能信息处理与人机交互技术等.

吴小培 男,(1966-),博士,教授,博士生导师. 研究方向为盲信号处理、医学信号处理等.