CNN在语音识别中的应用

总结目前语音识别的发展现状，dnn、rnn/lstm 和 cnn 算是语音识别中几个比较主流的方向。2012年，微软邓力和俞栋老师将前馈神经网络 FFDNN（Feed Forward Deep Neural Network）引入到声学模型建模中，将 FFDNN 的输出层概率用于替换之前 GMM-HMM 中使用 GMM 计算的输出概率，引领了 DNN-HMM 混合系统的风潮。长短时记忆网络（LSTM，LongShort Term Memory）可以说是目前语音识别应用最广泛的一种结构，这种网络能够对语音的长时相关性进行建模，从而提高识别正确率。双向 LSTM 网络可以获得更好的性能，但同时也存在训练复杂度高、解码时延高的问题，尤其在工业界的实时识别系统中很难应用。

回顾近一年语音识别的发展，deepcnn 绝对称得上是比较火的关键词，很多公司都在这方面投入了大量研究。其实 CNN 被用在语音识别中由来已久，在12、13 年的时候 OssamaAbdel-Hamid 就将 CNN 引入了语音识别中。那时候的卷积层和 pooling 层是交替出现的，并且卷积核的规模是比较大的，CNN 的层数也并不多，主要是用来对特征进行加工和处理，使其能更好的被用于 DNN 的分类。随着 CNN 在图像领域的发光发热，VGGNet，GoogleNet 和 ResNet 的应用，为 cnn 在语音识别提供了更多思路，比如多层卷积之后再接 pooling 层，减小卷积核的尺寸可以使得我们能够训练更深的、效果更好的 CNN 模型。

1 语音识别为什么要用 CNN

通常情况下，语音识别都是基于时频分析后的语音谱完成的，而其中语音时频谱是具有结构特点的。要想提高语音识别率，就是需要克服语音信号所面临各种各样的多样性，包括说话人的多样性(说话人自身、以及说话人间)，环境的多样性等。一个卷积神经网络提供在时间和空间上的平移不变性卷积，将卷积神经网络的思想应用到语音识别的声学建模中，则可以利用卷积的不变性来克服语音信号本身的多样性。从这个角度来看，则可以认为是将整个语音信号分析得到的时频谱当作一张图像一样来处理，采用图像中广泛应用的深层卷积网络对其进行识别。

从实用性上考虑，CNN 也比较容易实现大规模并行化运算。虽然在 CNN 卷积运算中涉及到很多小矩阵操作，运算很慢。不过对 CNN 的加速运算相对比较成熟，如 Chellapilla 等人提出一种技术可以把所有这些小矩阵转换成一个大矩阵的乘积。一些通用框架如 Tensorflow ， caffe 等也提供 CNN 的并行化加速，为 CNN 在语音识别中的尝试提供了可能。

下面将由“浅”入“深”的介绍一下 cnn 在语音识别中的应用。

2 CLDNN

提到 CNN 在语音识别中的应用，就不得不提 CLDNN（CONVOLUTIONAL, LONG SHORT-TERM MEMORY,FULLY CONNECTED DEEP NEURALNETWORKS）[1]，在 CLDNN 中有两层 CNN 的应用，算是浅层 CNN 应用的代表。CNN 和 LSTM 在语音识别任务中可以获得比 DNN

更好的性能提升，对建模能力来说，CNN 擅长减小频域变化，LSTM 可以提供长时记忆，所以在时域上有着广泛应用，而 DNN 适合将特征映射到独立空间。而在 CLDNN 中，作者将 CNN，LSTM 和 DNN 串起来融合到一个网络中，获得比单独网络更好的性能。

CLDNN 网络的通用结构是输入层是时域相关的特征，连接几层 CNN 来减小频域变化，CNN 的输出灌入几层 LSTM 来减小时域变化，LSTM 最后一层的输出输入到全连接 DNN 层，目的是将特征空间映射到更容易分类的输出层。之前也有将 CNN LSTM 和 DNN 融合在一起的尝试，不过一般是三个网络分别训练，最后再通过融合层融合在一起，而 CLDNN 是将三个网络同时训练。实验证明，如果 LSTM 输入更好的特征其性能将得到提高，受到启发，作者用 CNN 来减小频域上的变化使 LSTM 输入自适应性更强的特征，加入 DNN 增加隐层和输出层之间的深度获得更强的预测能力。

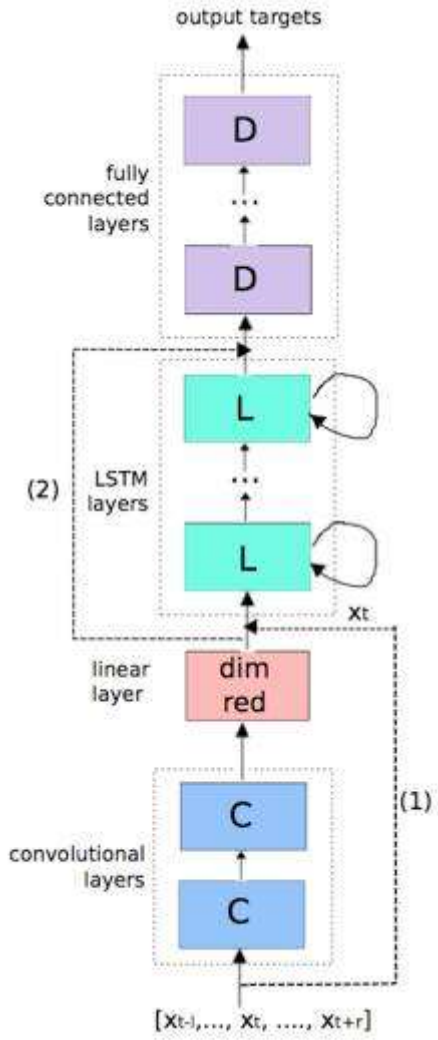

Fig 1. CLDNN Architecture

网络结构图如 图1 ，假设中心帧为

x_t

，考虑到内容相关性，向左扩展 L 帧，向右扩展 R 帧，则输入特征序列为

$[x_{t-l}, \ldots, x_{t+r}]$

，特征向量使用的是 40 维的 log 梅尔特征。

CNN 部分为两层 CNN，每层 256 个 feature maps，第一层采用 9x9 时域-频域滤波器，第二层为 4x3 的滤波器。池化层采用 max-pooling 策略，第一层 pooling size 是 3，第二层 CNN 不接池化层。

由于 CNN 最后一层输出维度很大，大小为 feature-maps*time*frequency，所以在 CNN 后 LSTM 之前接一个线性层来降维，而实验也证明降维减少参数并不会对准确率有太大影响，线性层输出为 256 维。

CNN 后接 2 层 LSTM，每个 LSTM 层采用 832 个 cells，512 维映射层来降维。

输出状态标签延迟 5 帧，此时 DNN 输出信息可以更好的预测当前帧。由于 CNN 的输入特征向左扩展了l帧向右扩展了 r 帧，为了确保 LSTM 不会看到未来多于 5 帧的内容，作者将 r 设为 0 。最后，在频域和时域建模之后，将 LSTM 的输出连接几层全连接 DNN 层。

借鉴了图像领域 CNN 的应用，作者也尝试了长短时特征，将 CNN 的输入特征

x_t

作为短时特征直接输入给 LSTM 作为部分输入，CNN 的输出特征直接作为 DNN 的部分输入特征。

3 deep CNN

在过去的一年中，语音识别取得了很大的突破。IBM、微软、百度等多家机构相继推出了自己的 Deep CNN 模型，提升了语音识别的准确率 Residual/Highway 网络的提出使我们可以把神经网络训练的更深。尝试 DeepCNN 的过程中，大致也分为两种策略：一种是 HMM 框架中基于 Deep CNN 结构的声学模型，CNN 可以是 VGG、Residual 连接的 CNN 网络结构、或是 CLDNN 结构。另一种是近两年非常火的端到端结构，比如在 CTC 框架中使用 CNN 或 CLDNN 实现端对端建模，或是最近提出的 Low Frame Rate、Chain 模型等粗粒度建模单元技术。

对于输入端，大体也分为两种：输入传统信号处理过的特征，采用不同的滤波器处理，然后进行左右或跳帧扩展。

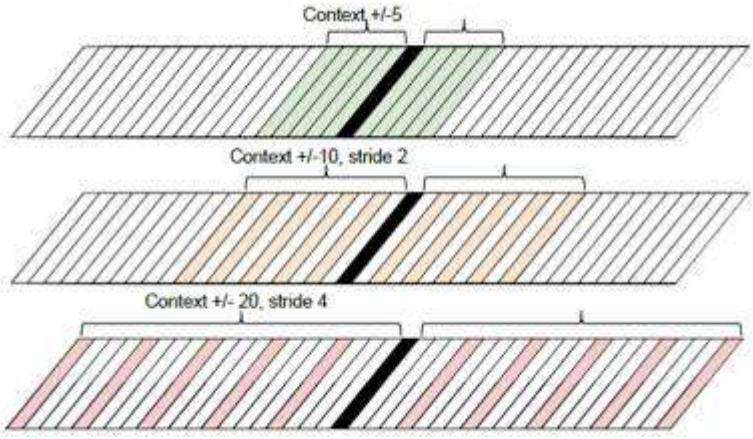

Fig 2.Multi-scale input feature. Stack 3*11*40

第二种是直接输入原始频谱，将频谱图当做图像处理。

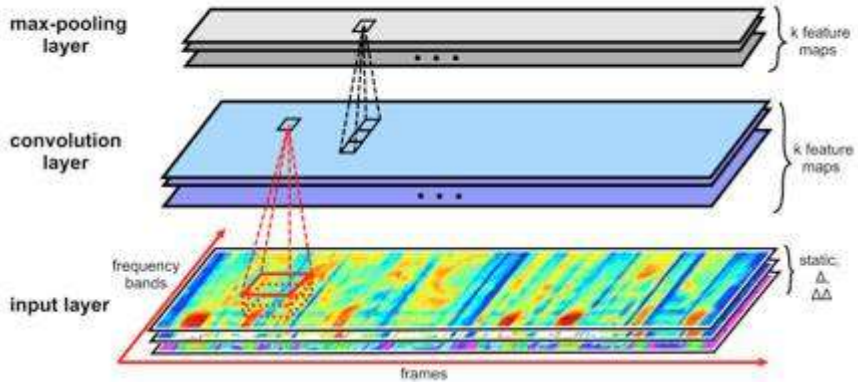

Fig 3. Frequency bands input

3.1 百度deep speech

百度将 Deep CNN 应用于语音识别研究，使用了 VGGNet，以及包含 Residual 连接的深层 CNN 等结构，并将 LSTM 和 CTC 的端对端语音识别技术相结合，使得识别错误率相对下降了 10% (原错误率的90%)以上。

此前，百度语音每年的模型算法都在不断更新，从 DNN ，到区分度模型，到 CTC 模型，再到如

今的 Deep CNN 。基于 LSTM-CTC 的声学模型也于 2015 年底已经在所有语音相关产品中得到了上线。比较重点的进展如下：

1)2013 年，基于美尔子带的 CNN 模型;

2)2014年，Sequence Discriminative Training(区分度模型);

3)2015 年初，基于 LSTM-HMM的语音识别；

4)2015 年底，基于 LSTM-CTC的端对端语音识别;

5)2016 年，Deep CNN 模型，目前百度正在基于 Deep CNN 开发 deep speech3，据说训练采用大数据，调参时有上万小时，做产品时甚至有10 万小时。

Fig4. 百度语音识别发展

百度发现，深层 CNN 结构，不仅能够显著提升 HMM 语音识别系统的性能，也能提升CTC 语音识别系统的性能。仅用深层 CNN 实现端对端建模，其性能相对较差，因此将如LSTM 或 GRU的循环隐层与 CNN结合是一个相对较好的选择。可以通过采用 VGG 结构中的 3*3 这种小 kernel ，也可以采用 Residual 连接等方式来提升其性能，而卷积神经网络的层数、滤波器个数等都会显著影响整个模型的建模能力，在不同规模的语音训练数据库上，百度需要采用不同规模的 DeepCNN 模型配置才能使得最终达到最优的性能。

因此，百度认为：

1)在模型结构中，DeepCNN 帮助模型具有很好的在时频域上的平移不变性，从而使得模型更加鲁棒(抗噪性);

2)在此基础上，DeepLSTM则与 CTC 一起专注于序列的分类，通过 LSTM 的循环连接结构来整合长时的信息。

3)在 DeepCNN 研究中，其卷积结构的时间轴上的感受野，以及滤波器的个数，针对不同规模的数据库训练的语音识别模型的性能起到了非常重要的作用。

4)为了在数万小时的语音数据库上训练一个最优的模型，则需要大量的模型超参的调优工作，依托多机多 GPU 的高性能计算平台，才得以完成工作。

5)基于 DeepCNN 的端对端语音识别引擎，也在一定程度上增加了模型的计算复杂度，通过百度自研的硬件，也使得这样的模型能够为广大语音识别用户服务。

3.2IBM

2015 年，IBM Watson 公布了英语会话语音识别领域的一个重大里程碑：系统在非常流行的评测基准 Switchboard 数据库中取得了 8% 的词错率（WER）。到了2016年 5 月份，IBM Watson 团队再次宣布在同样的任务中他们的系统创造了6.9% 的词错率新纪录，其解码部分采用的是 HMM，语言模型采用的是启发性的神经网络语言模型。声学模型主要包含三个不同的模型，分别是带有 maxout 激活的循环神经网络、3*3 卷积核的深度卷积神经网络、双向长短期记忆网络，下面我们来具体看看它们的内部结构。

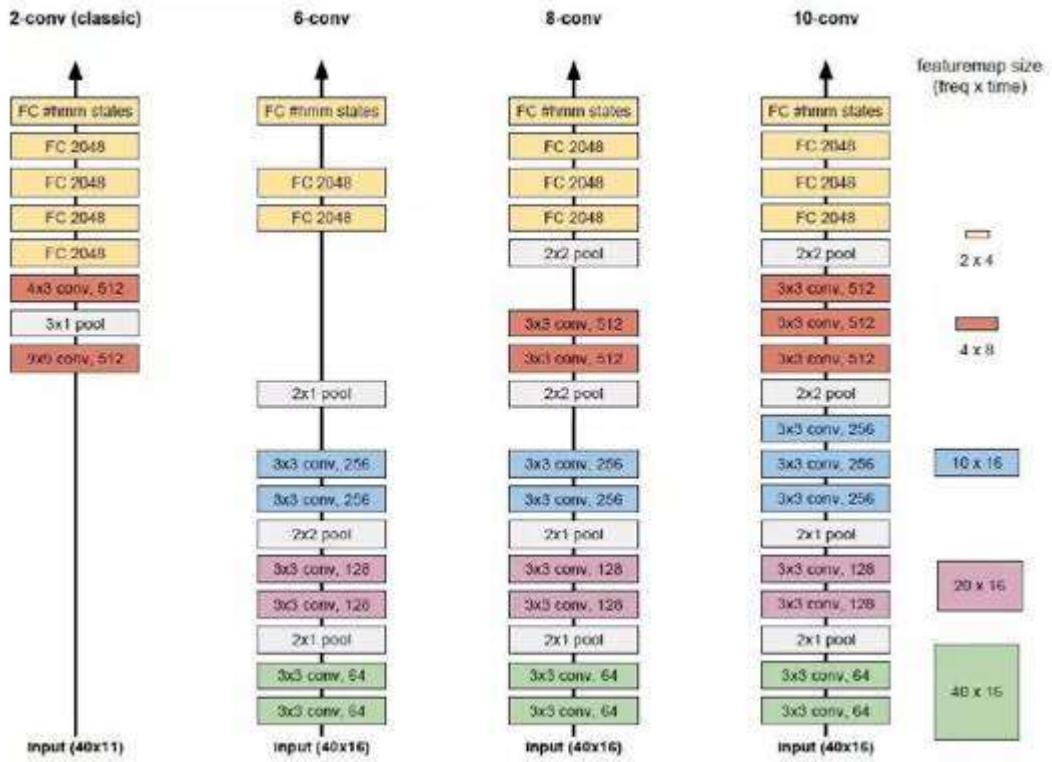

Fig 5. IBM Deep CNN 框架

非常深的卷积神经网络的灵感来自 2014ImageNet 参赛的 VGG 网络，中心思想是使用较小的 3*3 卷积核来取代较大的卷积核，通过在池化层之前叠加多层卷积网络，采取 ReLU 激活函数，可以获得相同的感知区域，同时具备参数数目较少和更多非线性的优点。

如上图所示，左1 为最经典的卷积神经网络，只使用了两个卷积层，并且之间包含一个池化层，卷积层的卷积核也较大，9*9 和 4*3，而卷积的特征面也较多，512 张卷积特征面。

左2、左3、左4均为深度卷积神经网络的结构，可以注意到与经典的卷积神经网络所不同的是，卷积的特征面由 64 个增加到 128 个再增加到 256 个，而且池化层是放在卷积的特征面数增加之前的；卷积核均使用的是较小的 3*3 卷积核，池化层的池化大小由 2*1 增加到 2*2 。

最右边 10-conv 的参数数目与最左边的经典卷积神经网络参数数目相同，但是收敛速度却足足快了 5 倍，尽管计算复杂度提高了一些。

3.3微软

2016年9月在产业标准 Switchboard 语音识别任务上，微软研究者取得了产业中最低的 6.3% 的词错率（WER）。基于神经网络的声学和语言模型的发展，数个声学模型的结合，把 ResNet 用到语音识别。

而在2016年的10月，微软人工智能与研究部门的团队报告出他们的语音识别系统实现了和专业速录员相当甚至更低的词错率（WER），达到了5.9%。5.9% 的词错率已经等同于人速记同样一段对话的水平，而且这是目前行Switchboard 语音识别任务中的最低记录。这个里程碑意味着，一台计算机在识别对话中的词上第一次能和人类做得一样好。系统性地使用了卷积和 LSTM 神经网络，并结合了一个全新的空间平滑方法（spatial smoothingmethod）和 lattice-free MMI 声学训练。

虽然在准确率的突破上都给出了数字基准，微软的研究更加学术，是在标准数据库——口语数据库 switchboard 上面完成的，这个数据库只有 2000 小时。

3.4 Google

根据 Mary Meeker 年度互联网报告，Google以机器学习为背景的语音识别系统，2017年3月已经获得英文领域95%的字准确率，此结果逼近人类语音识别的准确率。如果定量的分析的话，从2013年开始，Google系统已经提升了20%的性能。

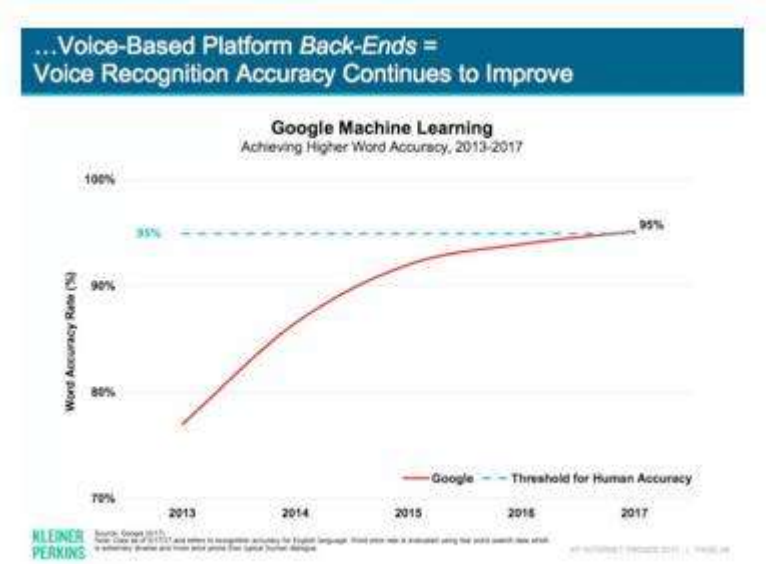

Fig 6. Google 语音识别性能发展

从近几年google在各类会议上的文章可以看出，google尝试deep CNN的路径主要采用多种方法和模型融合，如Network-in-Network(NiN)，Batch Normalization (BN)，ConvolutionalLSTM (ConvLSTM)方法的融合。比如2017 icassp会议中google所展示的结构

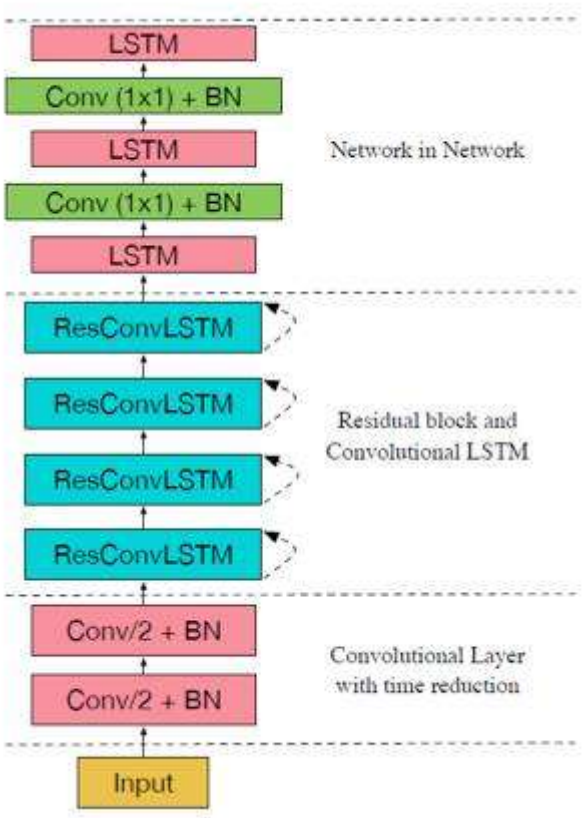

Fig 7. [5] includestwo convolutional layer at the bottom and followed by four residual block andLSTM NiN block. Each residual block contains one convolutional LSTM layer andone convolutional layer.

3.5科大讯飞DFCNN

2016年,在提出前馈型序列记忆网络FSMN (Feed-forward Sequential Memory Network) 的新框架后，科大讯飞又提出了一种名为深度全序列卷积神经网络（Deep Fully Convolutional Neural Network，DFCNN）的语音识别框架，使用大量的卷积层直接对整句语音信号进行建模，更好地表达了语音的长时相关性。

DFCNN的结构如下图所示，它输入的不光是频谱信号，更进一步的直接将一句语音转化成一张图像作为输入，即先对每帧语音进行傅里叶变换，再将时间和频率作为图像的两个维度，然后通过非常多的卷积层和池化（pooling）层的组合，对整句语音进行建模，输出单元直接与最终的识别结果比如音节或者汉字相对应。

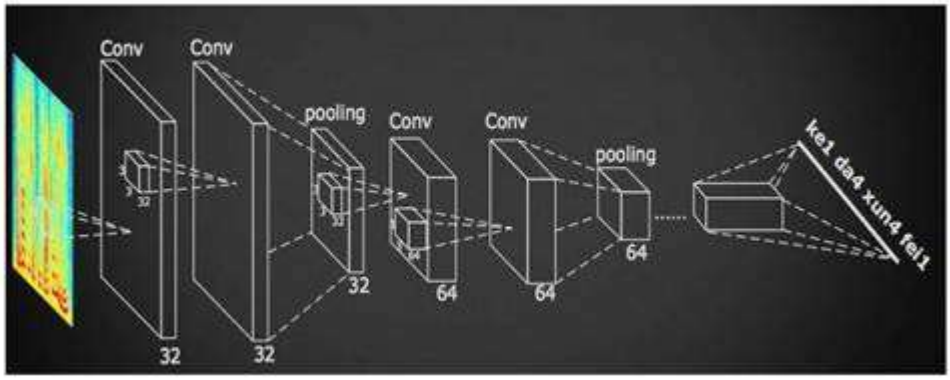

Fig 9. DFCNN框架

首先，从输入端来看，传统语音特征在傅里叶变换之后使用各种人工设计的滤波器组来提取特征，造成了频域上的信息损失，在高频区域的信息损失尤为明显，而且传统语音特征为了计算量的考虑必须采用非常大的帧移，无疑造成了时域上的信息损失，在说话人语速较快的时候表现得更为突出。因此DFCNN直接将语谱图作为输入，相比其他以传统语音特征作为输入的语音识别框架相比具有天然的优势。其次，从模型结构来看，DFCNN与传统语音识别中的CNN做法不同，它借鉴了图像识别中效果最好的网络配置，每个卷积层使用3x3的小卷积核，并在多个卷积层之后再加上池化层，这样大大增强了CNN的表达能力，与此同时，通过累积非常多的这种卷积池化层对，DFCNN可以看到非常长的历史和未来信息，这就保证了DFCNN可以出色地表达语音的长时相关性，相比RNN网络结构在鲁棒性上更加出色。最后，从输出端来看，DFCNN还可以和近期很热的CTC方案完美结合以实现整个模型的端到端训练，且其包含的池化层等特殊结构可以使得以上端到端训练变得更加稳定。

4总结

由于CNN本身卷积在频域上的平移不变性，同时VGG、残差网络等深度CNN网络的提出，给CNN带了新的新的发展，使CNN成为近两年语音识别最火的方向之一。用法也从最初的2-3层浅层网络发展到10层以上的深层网络，从HMM-CNN框架到端到端CTC框架，各个公司也在deep

CNN的应用上取得了令人瞩目的成绩。

总结一下，CNN发展的趋势大体为：

1 更加深和复杂的网络，CNN一般作为网络的前几层，可以理解为用CNN提取特征，后面接LSTM或DNN。同时结合多种机制，如attention model、ResNet 的技术等。

2 End to End的识别系统，采用端到端技术CTC， LFR 等。

3 粗粒度的建模单元，趋势为从state到phone到character，建模单元越来越大。

但CNN也有局限性，[2,3]研究表明，卷积神经网络在训练集或者数据差异性较小的任务上帮助最大，对于其他大多数任务，相对词错误率的下降一般只在2%到3%的范围内。不管怎么说，CNN作为语音识别重要的分支之一，都有着极大的研究价值。

[1] Sainath,T.N ,Vinyals, O., Senior, O.,Sak H:CONVOLUTIONAL, LONG SHORT-TERM MEMORY, FULLYCONNECTED DEEP NEURAL NETWORKS

[2] Sainath,T.N ，Mohamed,A.r ，Kingsbury ,B., Ramabhadran,B.:DEEPCONVOLUTIONAL NEURAL NETWORKS FOR LVCSR.In:Proc. International Conference onAcoustics, Speech and signal Processing(ICASSP),pp.8614-8618(2013)

[3] Deng, L.,Abdel-Hamid,O.,Yu,D.:ADEEP CONVOLUTIONAL NEURAL NETWORK USING HETEROGENEOUS POOLING FOR TRADINGACOUSTIC INVARIANCE WITH PHONETIC CONFUSION.In:Proc. International Conferenceon Acoustics, Speech and signal Processing(ICASSP),pp.6669-6673(2013)

[4] Chellapilla,K.,Puri, S., Simard,P.:High Performance Convolutional Neural Networks forDocument Processing.In: Tenth International Workshop on Frontiers inHandwriting Recognition(2006)

[5] Zhang, Y.,Chan ,W., Jaitly, N.:VERY DEEP CONVOLUTIONAL NETWORKS FOR END-TO-END SPEECHRECOGNITION.In:Proc. International Conference on Acoustics, Speech and signalProcessing(ICASSP 2017)