深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

语音识别：从入门到精通

第六讲：基于DNN-HMM的语音识别系统

主讲人 张彬彬

西北工业大学

binbzha@gmail.com

背景知识回顾（重要）

基于GMM-HMM的语音识别系统

- HMM通过状态跳转序列建模
- GMM概率密度建模
  - 每个状态都有各自独立的GMM
  - 识别时，计算在输入特征所有状态（GMM）上的概率打分，在哪个状态上概率最大。
  - 其实我们想做什么？分类，看当前输入属于哪个状态？

深度神经网络（Deep Neural Network）时代来了，怎么解决这个分类问题？

- 输入是什么？
- 输出是什么？
- 损失函数是什么？

注意

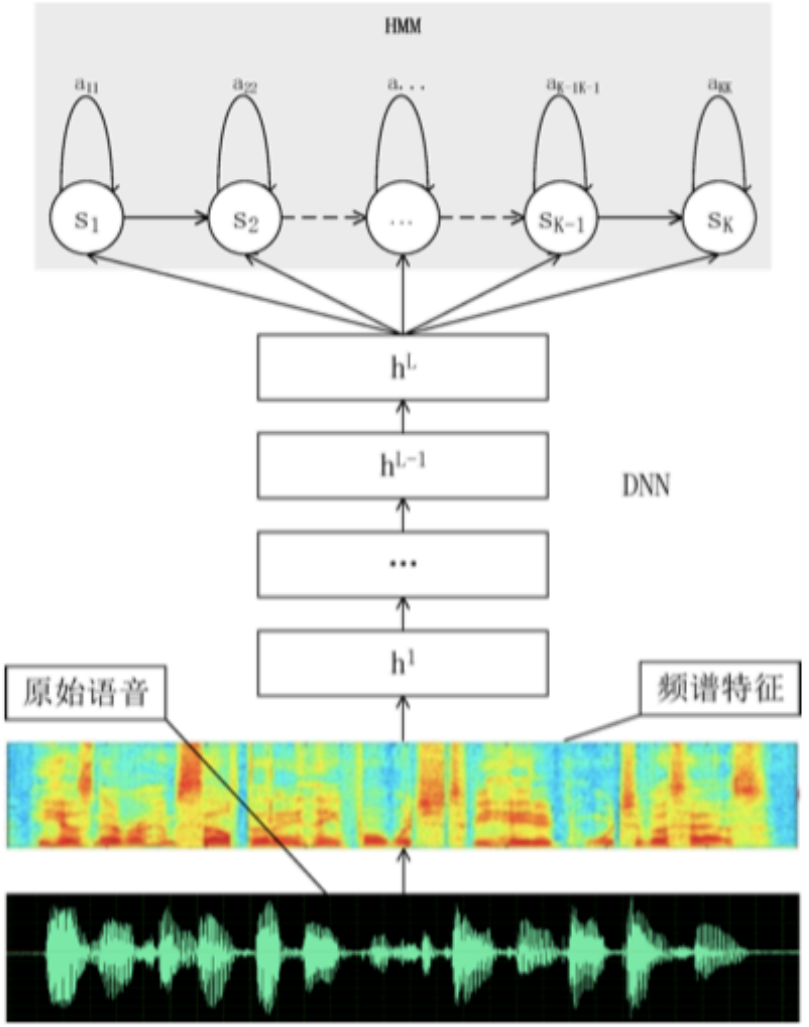
- 本节已假定读者已有一定的深度学习基础知识
- 本节目标
  - 回顾复习基本的深度神经网络知识
  - 重点带读者了解深度神经网络在语音识别中的应用
    - 成功应用的论文和时间
    - 带来了多少错误率的下降
- 所以，本文的重点是一些基本点，基本思想，并不会深入各种神经网络的细节。
- 此外，不是理所当然，本文中所述的每一种神经网络在语音识别中的成功应用，在当时都是里程碑式的贡献。

内容提要

- DNN-HMM语音识别系统
- 深度神经网络
  - 前馈神经网络FNN
  - 卷积神经网络CNN
    - CNN
    - TDNN
  - 循环神经网络RNN
    - LSTM
  - 混合神经网络
- 作业

DNN-HMM语音识别系统

- 输入是什么?
- 输出是什么?
- 损失函数是什么?



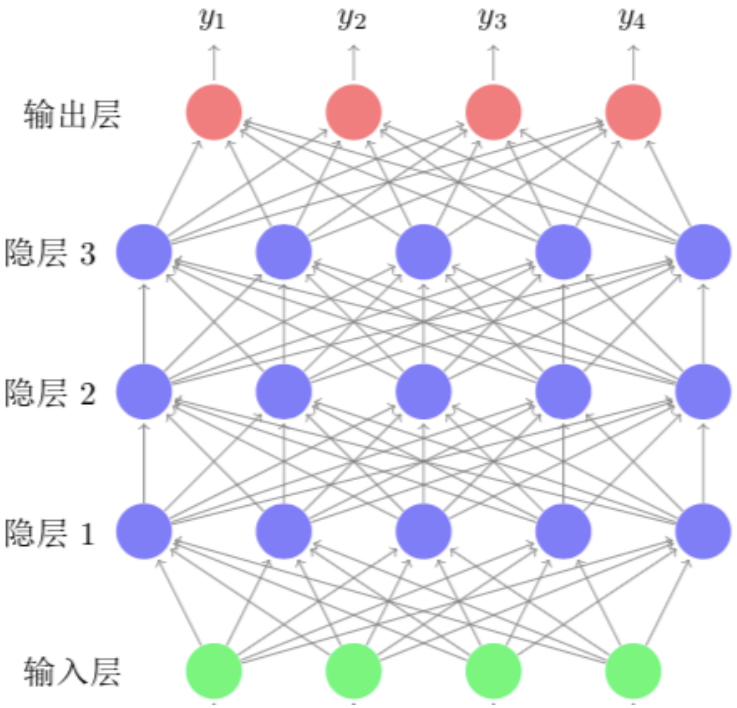
DNN分类问题损失函数

- Softmax概率归一化

$$y_k = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}$$

- 交叉熵CE(Cross Entropy)损失函数

$$L = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln(y_{nk})$$



GMM-HMM和DNN-HMM

GMM-HMM

- 每个状态有独立的GMM
- 模型输出概率：$\mathrm{p}(o_t|s_i)$

DNN-HMM

- 所有状态共享一个DNN
- 模型输出概率：$\mathrm{p}(s_i|\mathrm{o}_t)$
- HMM框架要的是什么？

$$p(o_t|s_i) = \frac{p(s_i|o_t)p(o_t)}{p(s_i)}$$

$$\begin{aligned}\log p(o_t|s_i) &= \log p(s_i|o_t) + \log p(o_t) - \log p(s_i) \\ &= \log p(s_i|o_t) - \log p(s_i)\end{aligned}$$

问题：DNN-HMM中如何得到状态的label?

DNN-HMM语音识别系统流程图

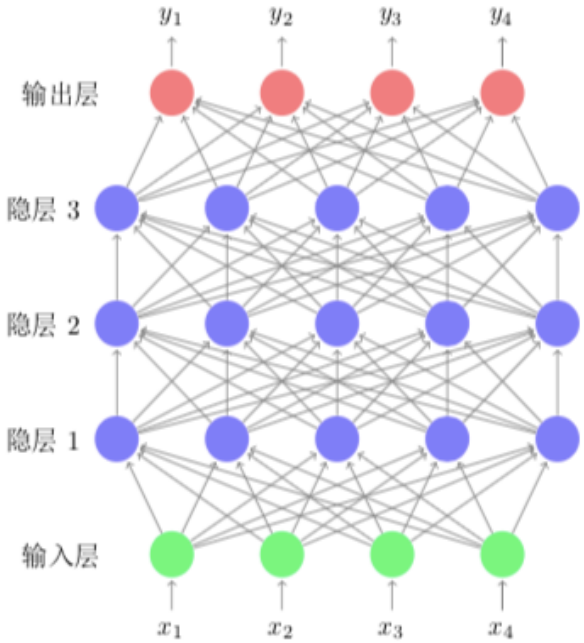
- 都要做哪些数据准备?
- 回想一下单音素训练过程?
- 再回想一下三音素训练过程?
- 何为DNN训练?

深度神经网络

- 网络类型
- 成功应用的论文和时间
- 错误率下降

FNN(Feedforward Neural Network)

$$\boldsymbol{y_l} = f(\boldsymbol{W_l x} + \boldsymbol{b_l})$$



激活函数

- sigmoid

$$s(z) = \frac{1}{1 + e^{-z}}$$

- tanh

$$\tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

- ReLU(Rectified Linear Unit)

$$\mathrm{Re}LU(z) = \max(0, z)$$

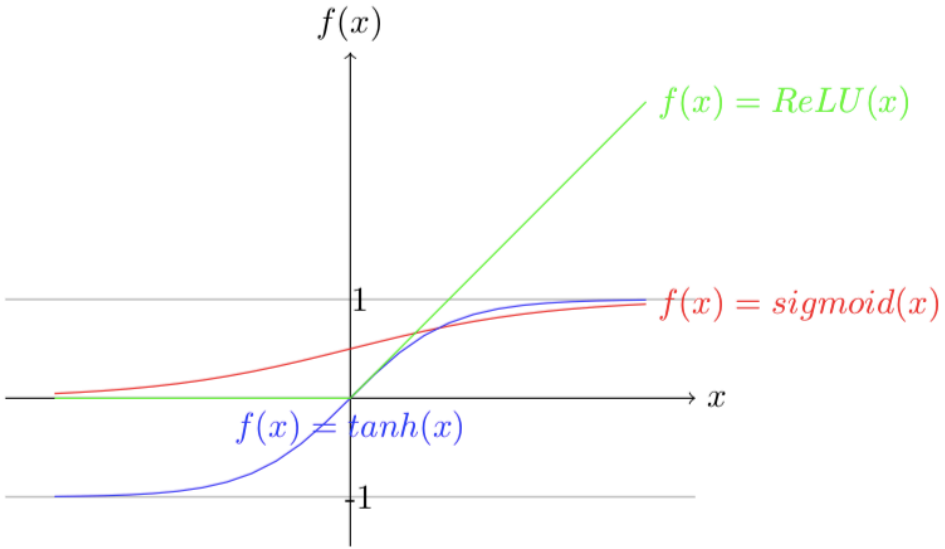


图 3-3 激活函数 *sigmoid*、*tanh* 和 *ReLU* 对比

FNN在语音识别中的应用

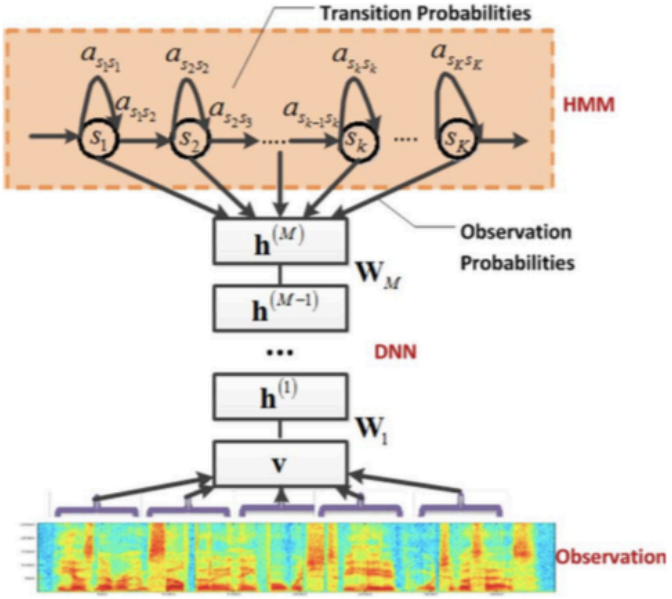


Fig. 1. Diagram of our hybrid architecture employing a deep neural network. The HMM models the sequential property of the speech signal, and the DNN models the scaled observation likelihood of all the senones (tied tri-phone states). The same DNN is replicated over different points in time.

TABLE II
CD-GMM-HMM BASELINE RESULTS

Criterion	Dev Accuracy	Test Accuracy
ML	62.9%	60.4%
MMI	65.1%	62.8%
MPE	65.5%	63.8%

TABLE VI
EFFECTS OF ALIGNMENT AND TRANSITION PROBABILITY TUNING ON BEST DNN ARCHITECTURE

Alignment	Tune Trans.	Dev Acc	Test Acc
from CD-GMM-HMM ML	no	70.3%	68.4%
from CD-GMM-HMM MPE	no	70.7%	68.8%
from CD-GMM-HMM MPE	yes	71.0%	69.0%
from CD-DNN-HMM	no	71.7%	69.6%
from CD-DNN-HMM	yes	71.8%	69.6%

- 错误率GMM->DNN
  - Dev: 37.1% -> 28.1%
  - Test: 39.6% -> 31.4%
- 错误率下降 20%

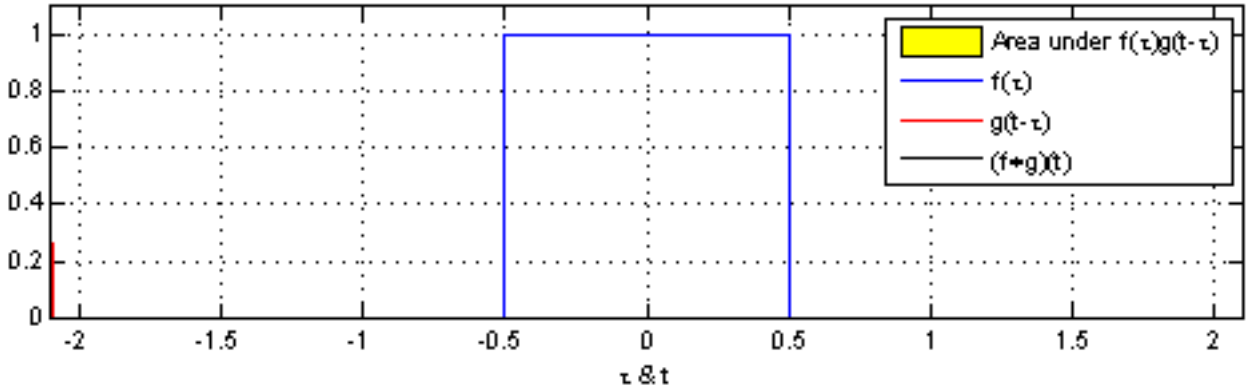
2012 - Dahl et al. - Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

CNN(Convolution Neural Network)

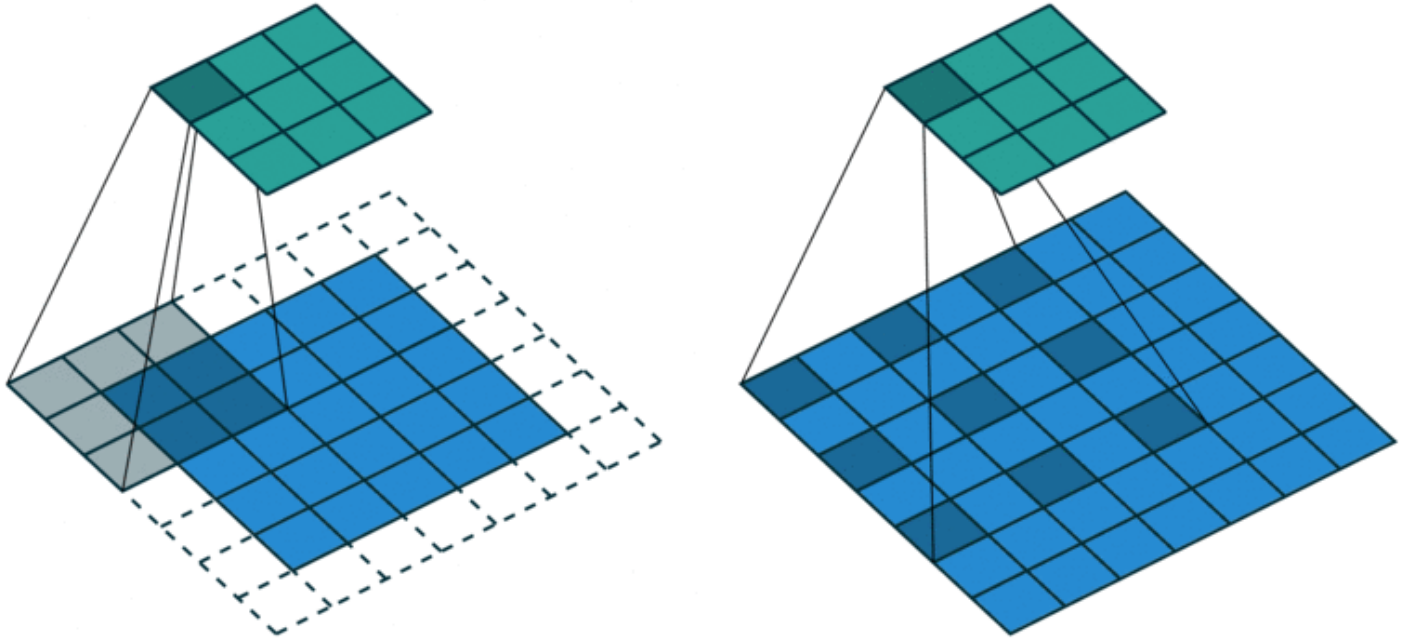
- Convolution(卷积，信号处理)

$$(f * g)[t] = \sum_{\tau} f(\tau) g(t - \tau)$$

- 平移
- 点乘
- 求和



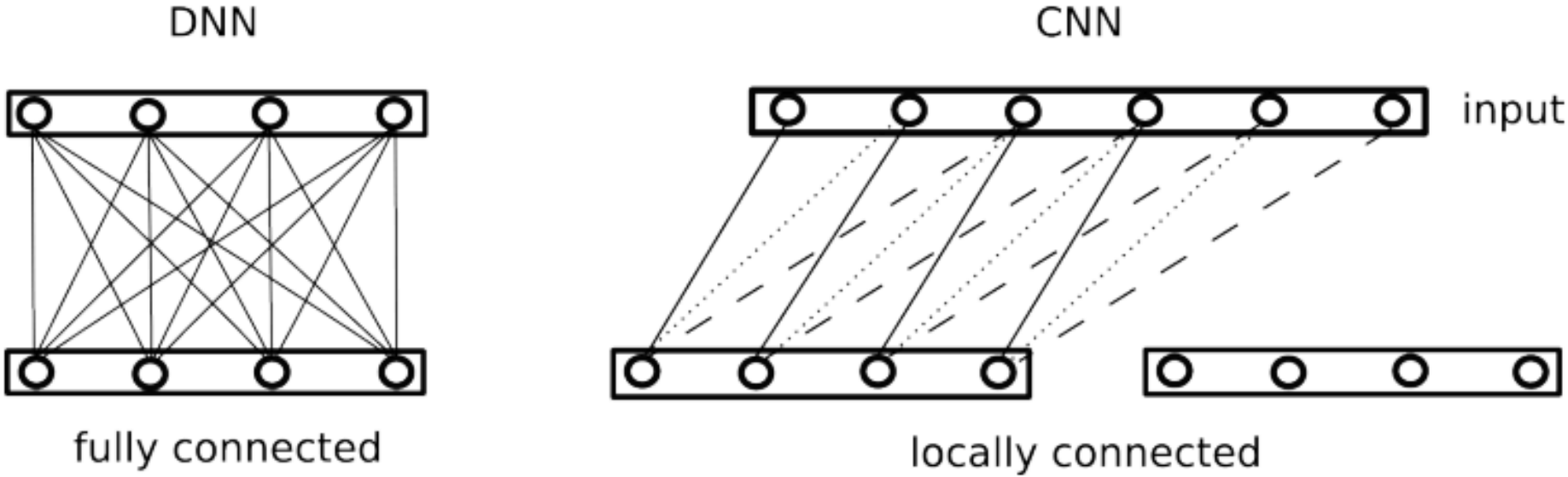
CNN



CNN

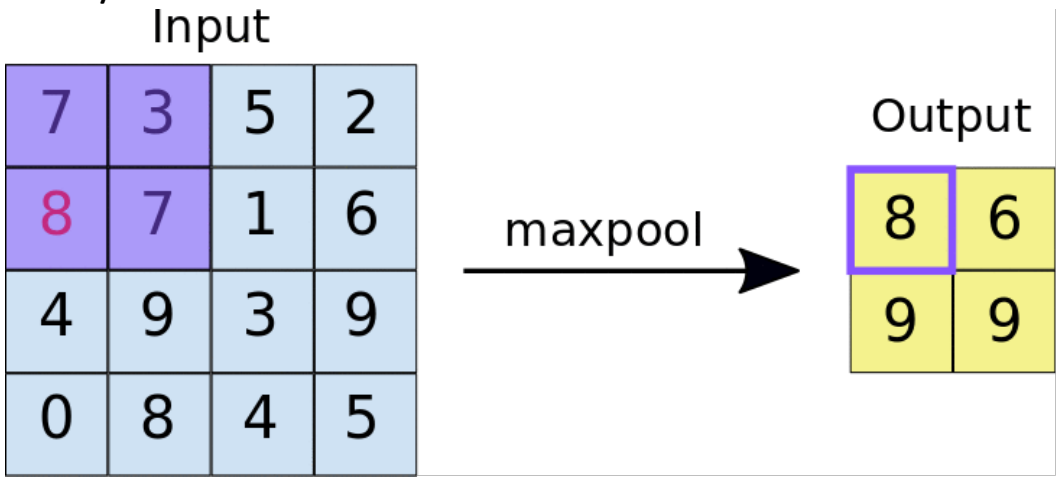
$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$$

CNN/DNN



Pooling

- 类型
  - Max Pooling
  - Average Pooling
- 作用
  - Dimension Reduction
  - Invariance/Robust



CNN在语音识别中的应用-CNN

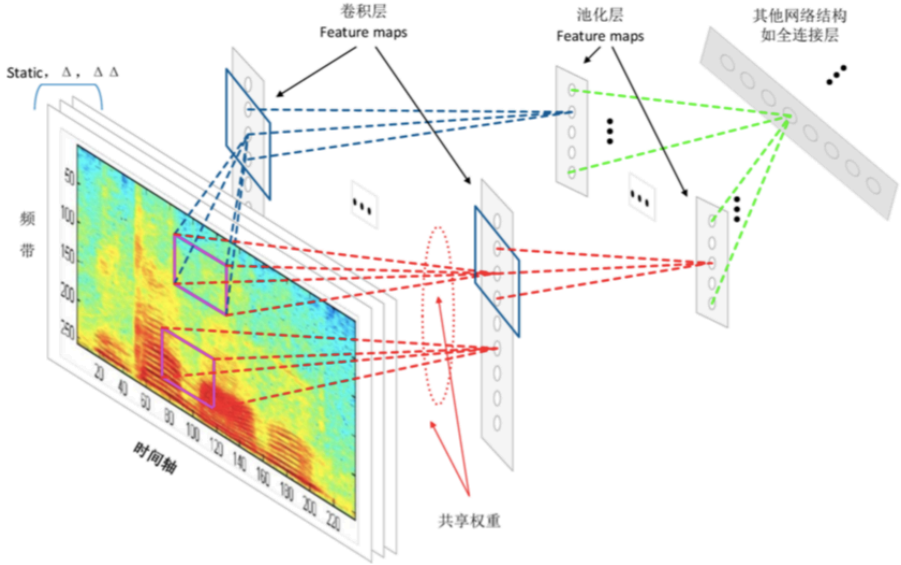


图 3-5 基于 CNN 的声学模型

Table 1: *Comparisons of TIMIT phone recognition accuracy among different CNN architectures. LWS: limited weight sharing; FWS: full weight sharing; K: # of feature maps; PS: pooling size; FS: filter size; B: # of bands.*

Convolution architecture	PER
No convolution	22.9 %
Freq FWS (K:200, PS:6, FS:8, B:20)	21.6%
Freq LWS (K:84, PS:6, FS:8, B:20)	20.5%
Time FWS (K:400, PS:2, FS:8, B:7)	22.5%
2D Multi-layers (K:40, PS:2,2, FS:3,3, B:20,7), (K:200, PS:3,1, FS:5,7, B:18,1)	21.5%

- 5~10%错误率下降
- 2D CNN在当时没有取得比较好的效果

2013 - Abdel-Hamid, Deng, Yu - Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition

CNN在语音识别中的应用-TDNN

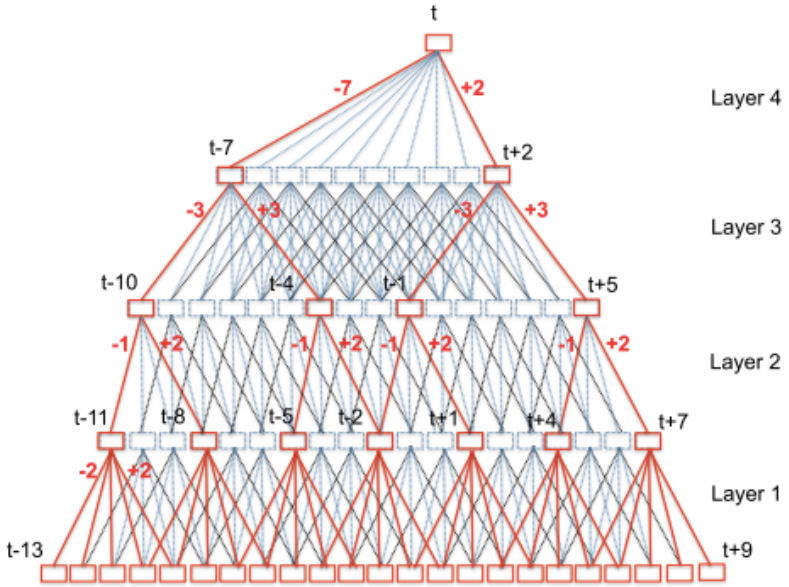


Figure 1: Computation in TDNN with sub-sampling (red) and without sub-sampling (blue+red)

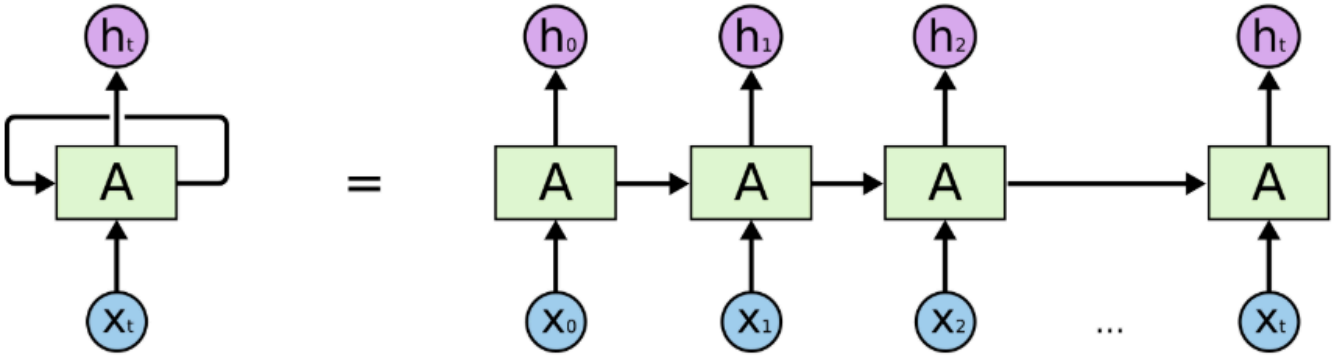
Table 4: Baseline vs TDNN on various LVCSR tasks with different amount of training data

Database	Size	WER		Rel.
		DNN	TDNN	Change
Res. Management	3h hrs	2.27	2.30	-1.3
Wall Street Journal	80 hrs	6.57	6.22	5.3
Tedlium	118 hrs	19.3	17.9	7.2
Switchboard	300 hrs	15.5	14.0	9.6
Librispeech	960 hrs	5.19	4.83	6.9
Fisher English	1800 hrs	22.24	21.03	5.4

- TDNN(Time Delay Neural Network)和扩张卷积的想法是一致的
- 仅时域卷积，没有pooling
- 5~10%错误率下降

2015 - Peddinti, Povey, Khudanpur - A time delay neural network architecture for efficient modeling of long temporal contexts

RNN(Recurrent Neural Network)

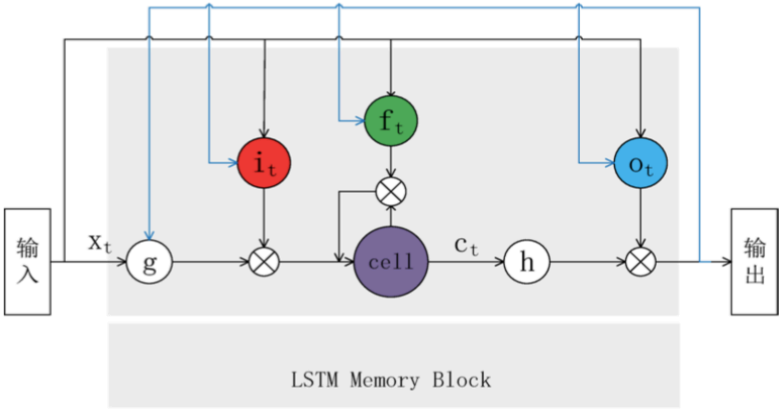


An unrolled recurrent neural network.

$$\mathbf{h}_t = f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1})$$

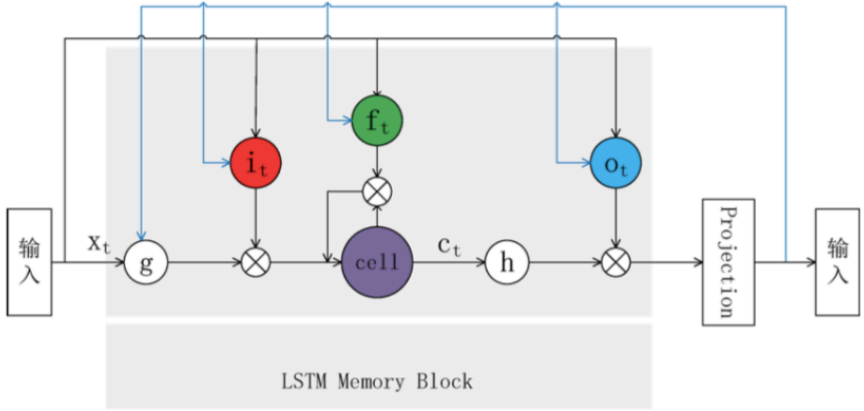
其中 $f(.)$ 表示激活函数, $\mathbf{W}_{xh}$ 是 $N \times M$ 的连接前一层的权值矩阵， $\mathbf{W}_{hh}$ 是 $N \times N$ 的连接 $t-1$ 时刻该循环层输出 $\mathbf{h}_{t-1}$ 的权值矩阵， $\mathbf{h}_{t-1}$ 即是 RNN 的内部状态。

LSTM (Long Short Term Memory)



a) LSTM

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c)$$
$$h_t = o_t \odot \tanh(c_t)$$



b) LSTMP

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c)$$
$$m_t = o_t \odot \tanh(c_t)$$
$$h_t = W_m m_t$$

LSTM在语音识别中的应用

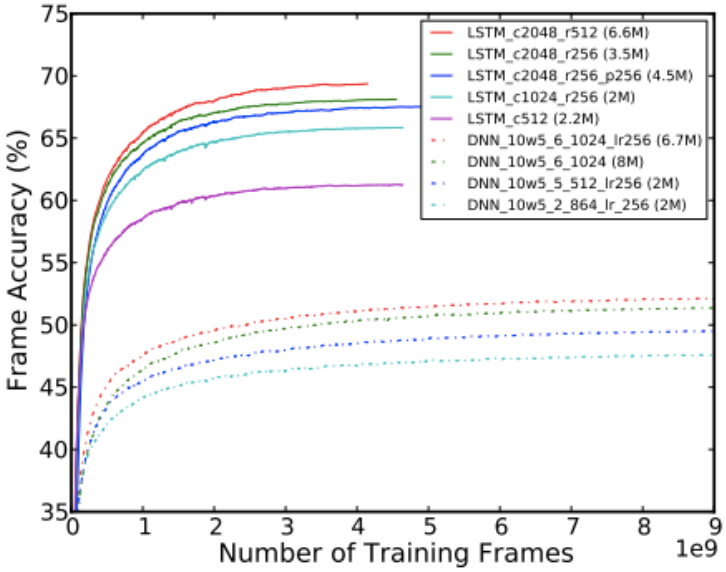


Fig. 3. 2000 context dependent phone HMM states.

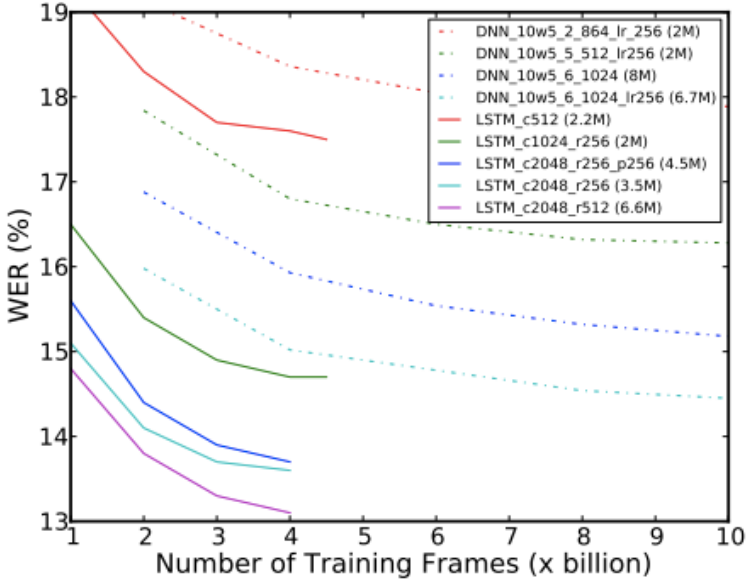


Fig. 6. 2000 context dependent phone HMM states.

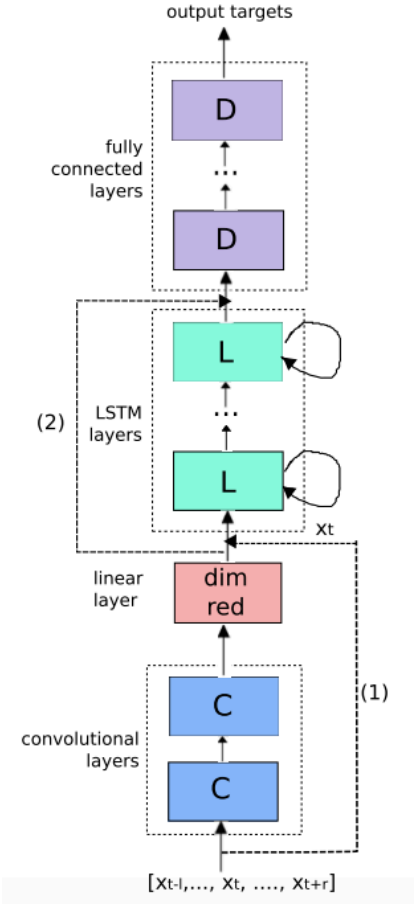
- 训练帧正确率高很多
- 5~10%错误率下降

2014 - Sak, Senior, Beaufays - Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition

混合神经网络

- FNN
  - 全局特征抽取
- CNN
  - 局部特征抽取
  - Invariance
- RNN
  - 记忆
  - 时序建模能力
- 复杂网络基本是以上三种网络的组合

CLDNN



Method	WER-CE	WER-Seq
LSTM	20.3	18.8
CLDNN	19.4	**17.4**
multi-scale CLDNN	**19.2**	**17.4**

Table 9. WER, Models Trained on 2,000 hours, Noisy

2015 - Sainath et al. - Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks

其他NN知识

- BP/BPTT
- Optimizer(SGD/Momentum/Adam …)
- Dropout
- Regularization
- Residual Connection
- Batch Normalization

本章总结

- DNN-HMM语音识别系统
- 深度神经网络
  - 前馈神经网络FNN
  - 卷积神经网络CNN
    - CNN
    - TDNN
  - 循环神经网络RNN
    - LSTM
  - 混合神经网络
- 作业

作业1

- 作业地址 https://github.com/nwpuaslp/ASR_Course/tree/master/06-DNN-HMM
- 作业1：完善DNN代码，并基于该DNN实现11个数字识别
  - 基本实验：拓展ReLU和FullyConnect的前向后向算法
  - 拓展1: 超参数如学习率、隐层数、隐层节点数
  - 拓展2: 基于该框架实现神经网络的一些基本算法，如
    - sigmoid和tanh激活函数
    - dropout
    - L2 regularization
    - optimizer(momentum/adam)
- 作业2：基于Kaldi和THCHS30理解梳理基于DNN-HMM的语音识别系统。
  - 基本流程步骤
  - 每一步骤的输入、输出
  - 步骤间的依赖关系

感谢各位聆听！