基于 DNN 与 RNN 声学模型融和的语音识别研究

朱会峰，贺勇，雍坤，丁沛，郝杰
东芝（中国）有限公司研究开发中心 北京 100600 中国

文 摘：采用上下文依赖的神经网络隐马尔科夫模型（CD-NN-HMM）的语音识别系统在商业领域已经成为主流，其性能已经被验证大幅度超过传统的高斯混合模型（GMM）。其中深度神经网络（DNN）与递归神经网络（RNN）是 CD-NN-HMM 架构中两种最具代表性的模型。本文研究了将 DNN 与 RNN 分别在拓扑结构与声学得分层面进行融合的新型语音识别方法。实验结果表明 DNN 与 RNN 声学模型融合的方法可以有效的提高语音识别精度。

关键词：DNN；RNN；语音识别；声学模型融合
中图分类号： TN912.34

先进机器学习算法和智能快速计算设备使语音识别性能在最近几年快速提高。在基于 GMM-HMM 的语音识别框架中，首先通过前向后向算法将句子对齐，然后使用期望最大化（EM）算法迭代训练高斯混合模型（GMM）[1]，在识别阶段以 GMM 输出作为 HMM 状态的概率输出得分。GMM-HMM 曾经是商业大规模连续语音识别系统的主流。近几年，由于计算机硬件设备 GPU 的发展，在大规模数据上训练深层次的人工神经网络的时间大大缩短，使得基于上下文依赖的神经网络隐马尔科夫模型（CD-NN-HMM）[2]广泛应用到语音识别领域，在识别性能上大幅度的超过了 GMM-HMM。

多伦多大学 2009 年发表文章首次提出基于深度信念网络的深度神经网络(DBN-DNN)的声学模型，在 TIMIT 数据集上识别性能超过了 GMM 模型[3]，微软首次将 DBN-DNN 的声学模型应用到大规模连续语音识别的商业系统中，结果显示识别精度的提高超过 30%[4]，谷歌，IBM 的实验结果均显示 DNN 使得语音识别性能大幅度提升[5][6]。DBN-DNN 的声学模型首先基于 DBN 的无监督的预训练帮助 DNN 在参数空间中找到一个更好的搜索起点。微软使用逐层扩展的有监督的预训练也有类似的效果[7]，然后使用随机梯度下降算法寻找最优点。DNN 与普通人工神经网络的区别是 DNN 有更深的层数，使得 DNN 可以更加有效的学习和提取数据的深层信息，从而提高分类能力。

递归神经网络 RNN 与 DNN 这种前向神经网络不同，其是一种带有反馈链接的神经网络。在时间这个维度上，RNN 是有深度的，相对于固定时间维度的 DNN，RNN 可以记录更长的历史信息。但是

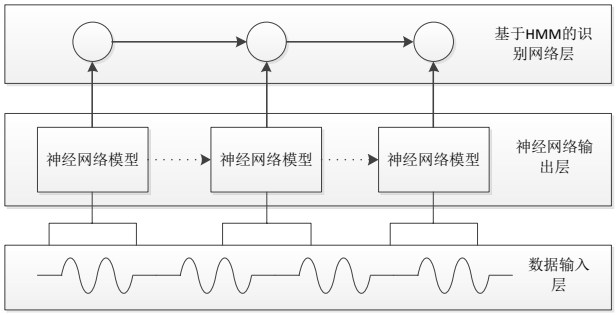


图 1 CD-NN-HMM 的语音识别框架

由于梯度消失的问题，传统的 RNN 不能记录较长的历史信息，为了解决这个问题，慕尼黑工业大学 1997 年提出了一种复杂的 RNN 结构长短时记忆模型（LSTM）[11]，其在基于序列的模式识别中表现出良好的性能。2013 年，多伦多大学使用这种复杂的 RNN 结构 LSTM 在 TIMIT 数据集上取得目前最好的识别结果[8][9]，谷歌则采用分布式的训练系统在 1900 小时的语音数据上基于 CD-NN-HMM 的框架上训练一种改进的 LSTM 网络结构，实验结果显示其性能超过了 DNN[10]。

在图 1 的 CD-NN-HMM 的识别框架中，神经网络模型可以选择不同拓扑结构，其中 DNN 与 RNN 是最具代表性的两种模型。在这个框架中，DNN 与 RNN 的区别在于 DNN 在神经网络层的输出不依赖前一个时间点的输出，RNN 则依赖前一个时间点的输出。本文基于 CD-NN-HMM 的框架，研究了将 DNN 与 RNN 分别在拓扑结构与声学得分层面进行融合的语音识别方法，实验表明该方法可以有效的提高语音识别精度。

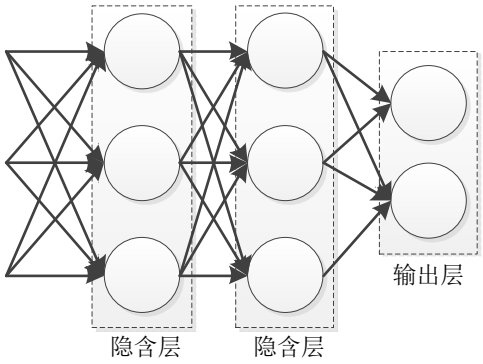


图 2 前向神经网络结构

1 神经网络结构

1.1 深度神经网络

深度神经网络是具有多个隐含层的前向神经网络。 它的拓扑结构如图 2 所示，对于每个链接都有一个权重，层中每个节点是最小的计算单元，每个节点的输出作为下一层节点的输入，每个节点的计算公式如下

$$a_k = \sum_{i=1}^{i=n} w_{ki} x_i \tag{1}$$

$$y_k = \begin{cases} f(a_k) & k\text{是隐含层节点} \\ g(a_1, \dots, a_k, \dots, a_m) & k\text{是输出层节点} \end{cases} \tag{2}$$

其中公式(1)表示在节点 k 对 n 个输入做线性的加和，(2)表示对节点线性加和做非线性的函数映射，对于隐含层常用的如 sigmoid，tanh，Relu 等，输出层常用的如 softmax。

1.2 递归神经网络

递归网络是带有反馈连接的神经网络，它依时序的输入做相应的输出，本文的实验是基于改进的 LSTM 递归神经网络[10]，它的拓扑结构如图 3 所示，图 3 中 LSTM 层的节点是文章[12]提出的记忆单元，这个记忆单元除了传统的 RNN 中的时间序列上的反馈输入，内部还有输入因子控制单元、输出因子控制单元和自记忆控制单元。LSTM 记忆单元有效的克服了梯度消失的问题，从而可以记忆更长的历史信息。反馈层的节点是普通人工神经网络的节点，计算公式参考深度神经网络中的公式(1)(2)。

2 DNN 与 RNN 声学模型融合方法

2.1 拓扑结构融合

拓扑结构融合方法如图 4 所示，底层使用改进的 LSTM-RNN 递归的神经网络结构，在高层使用 DNN 的神经网络结构。因为 LSTM-RNN 中记忆单元的改进，可以记录更长的历史信息，而 DNN 作为深度神经网络可以更加有效的提取数据中高层次信息的特征，从而提高分类精度，因此本文提出在底层使用 LSTM-RNN 去提取记录更长历史信息的数据特征表示，然后使用深层次的神经网络提取更加有效的数据特征，从而可以结合 RNN 与 DNN 的优点。使用拓扑结构融合的方法去改善识别率是在图 1 的 CD-NN-HMM 框架中使用图 4 的网络结构计算 HMM 模型中的状态中输出概率得分

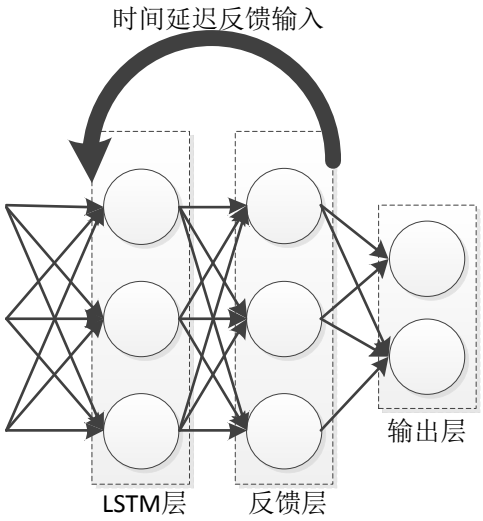


图 3 改进的 LSTM 递归网络

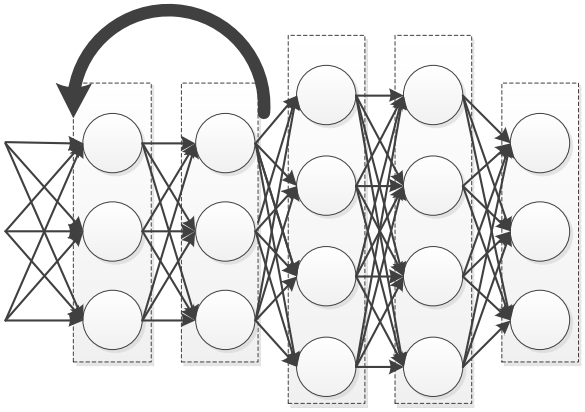


图 4 DNN 与改进的 LSTM-RNN 的融合结构

2.2 声学得分融合

在 CD-NN-HMM 的框架中 HMM 状态概率输出得分计算公式如下

$$\log p(o|s) = \log p(s|o) - \alpha \log p(s) \tag{3}$$

其中 o 表示特征向量，s 是 HMM 模型中的状态，α是调节先验概率影响的因子，p(s|o)是神经网络的输出层对应状态的后验概率值。

概率输出得分的融合公式是基于logp(o|s)在 DNN 与 RNN 的得分上作插值，从而得到一个更加平滑的状态概率输出得分可由下式表示：

$$\log p(o|s) = \varphi \log p(o|s)_{dnn} + (1-\varphi) log p(o|s)_{rnn} \quad (4)$$

其中φ是插值调节因子。

3 实验及结果分析

3.1 模型的训练与评测信息

本文使用最小化交叉熵准则训练神经网络模型。深度神经网络使用 BP 算法，改进的 LSTM 的递归神经网络使用 BPTT 算法。训练策略是每次迭代结束后，在交叉验证集合验证，如果性能提升则记录最好模型，否则当前学习率乘以缩放因子，如果减少的学习率低于最小学习率则取最小学习率，在尝试 4 次没有得到更好的训练模型时，训练结束。

声学模型训练的数据规模是 40 小时普通话语音数据，交叉验证的数据规模是 4 个小时，识别率的评测是在 475 句话的数据上评测。实验中采用的是基于 CD-NN-HMM 的 WFST 的解码器，语言模型是 4 元的 ARPA 模型。声学特征使用倒谱加基频的 74 维特征数据，使用 GMM 模型将特征数据做对齐得到标签数据。

3.2 拓扑结构融合实验

表 1.结构融合实验结果

神经网络拓扑结构	训练集准确率(%)	验证集准确率(%)	测试集识别率(%)
814:(512H)*3:4510(DNN1)	58.2	53.9	84.71
74:400L:256R:4510(RNN1)	63.1	58.0	82.94
RNN1⊕DNN1(MIX1)	68.5	62.3	83.66
814:(1024H)*4:4510(DNN2)	65.3	57.2	85.07
74:800L:512R:4510(RNN2)	69.0	61.4	82.67
RNN2⊕DNN2（MIX2）	67.3	60.8	83.02

表 1 中 (512H)*3 表示 3 个 512 节点的隐含层，400L 表示一个有 400 个计算节点的 LSTM 层，256R 表示一个有 256 个节点的反馈层,RNN1⊕DNN1 表示图 4 中的融合的神经网络结构。DNN 的输入是一个前后加 5 桢的 11 桢固定窗口，其维数为 814，改进的 LSTM-RNN 和融合的神经网络结构的输入节点个数是当前帧维数 74。

表 1 中实验显示 LSTM-RNN 对训练集合验证集合的分类准确率远高于 DNN，扩大神经网络结构可以提高分类准确率，但是在识别精度上，DNN(DNN1,DNN2) 表现的优于改进的 LSTM-RNN(RNN1,RNN2)与融合的神经网络结构(MIX1,MIX2)，对于改进的 LSTM-RNN 来说，高分类准确率但是低识别精度的结果说明，在小数据训练规模上，改进后的 LSTM-RNN 的泛化性不如

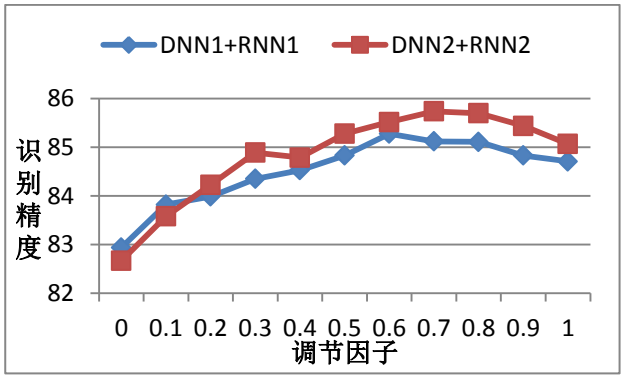
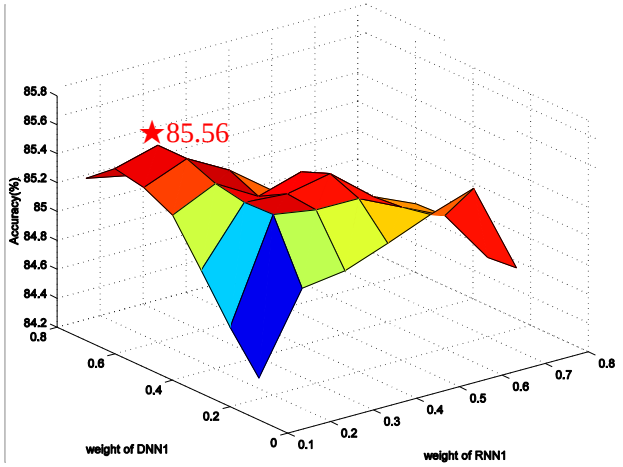
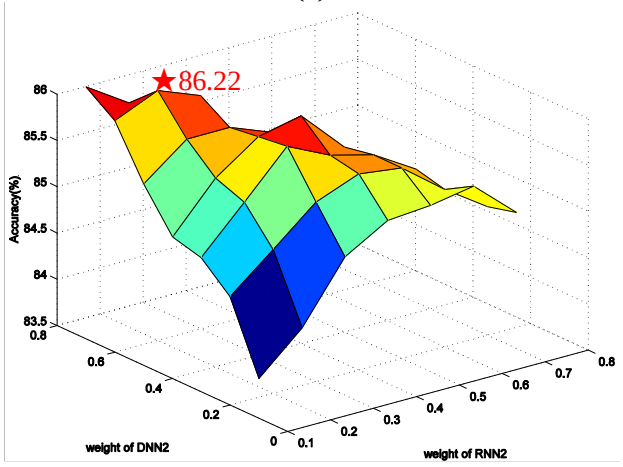


图 5 DNN 与 RNN 声学得分插值实验结果



(a)



(b)

图 6 DNN，RNN 与 MIX 声学得分插值实验结果

DNN。RNN1 与 RNN2 的结果对比显示对于复杂的模型，小数据规模上扩大其模型结构，容易陷入过拟合。融合后的神经网络结构（MIX1，MIX2）相对于改进后的 LSTM-RNN（RNN1,RNN2），识别精度明显提高，说明融合的神经网络结构的泛化性优于改进后的 LSTM-RNN。

3.2 声学得分融合实验

图 5 中，根据公式（4）计算概率输出得分，调整参数φ从 0 到 1，步长为 0.1，分别在 DNN1 与 RNN1，DNN2 与 RNN2 上做了声学得分插值实验。对于 DNN1 与 RNN1 的声学得分插值实验，在φ=0.6 的

时候取得最优识别精度 85.28%，相对于 DNN1 的结果 84.71%, 相对误识率降低了 5.7%。对于 DNN2 与 RNN2 的声学得分插值实验，在φ=0.7 的时候取得最优识别精度 85.74%，相对于 DNN2 的结果 85.07%，相对误识率降低 4.5%。图 6 中进一步对 DNN,改进后的 LSTM-RNN 以及融合的神经网络这 3 个模型做声学得分插值融合，图 a 显示 DNN1，RNN1 与 MIX1 的插值融合可以进一步把识别精度提高到 85.56%，相对于 DNN1 与 RNN1 插值融合的结果 85.28%, 相对误识率降低了 1.9%，相对于 DNN1 的结果 84.71%,相对误识率降低 5.6%。同样图 b 中 DNN2，RNN2 与 MIX2 这 3 个模型的插值融合结果 86.22%相对于 DNN2 与 RNN2 插值融合结果 85.74%，相对误识率降低了 3.4%,相对于 DNN2 的结果 85.07%相对误识率降低 7.7%。图 5 与图 6 的实验结果均说明了不同类型的神经网络做声学得分融合，其性能表现优于单个神经网络，明显的降低了识别错误率。

4 结论

本文在基于 CD-NN-HMM 的语音识别框架下，提出在网络结构和概率输出得分上融合 DNN 与 RNN 的声学模型方法，在小规模训练数据上验证了底层使用改进的 LSTM-RNN 的网络结构，高层使用 DNN 的网络结构可以有效的提高融合后的 LSTM-RNN 的泛化性。对不同类型的神经网络在声学得分上作插值融合，可以有效的平滑 HMM 中状态的概率输出，提高识别精度。

参 考 文 献

[1] B. H. Juang, S. Levinson, and M. Sondhi, “Maximum likelihood estimation for multivariate mixture observations of Markov chains[J],” IEEE Transactions on Information Theory, vol. 32, no. 2, pp. 307–309, 1986

[2] H. Bourlard and N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, Norwell, MA, USA, 1993

[3] A. Mohamed, G. Dahl, and G. Hinton. “Deep belief networks for phone recognition,”[A] in NIPS Workshop on Deep Learning for Speech Recognition and Related Applications[C], 2009

[4] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,”[A] in Proc. Interspeech 2011[C], pp. 437–440

[5] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, “An Application of Pretrained Deep Neural Networks To Large Vocabulary Conversational Speech Recognition,”[R] Tech. Rep. 001, Department of Computer Science, University of Toronto, 2012

[6] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, “Improvements in using deep belief networks for large vocabulary continuous speech recognition,”[R] Tech. Rep. UTML TR 2010-003, Technical Report, Speech and Language Algorithm Group, IBM, February 2011

[7] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,”[A] in Proc. IEEE ASRU[C], 2011, pp. 24–29.

[8] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,”[A] in Proc. ICASSP[C], Vancouver, Canada, 2013, pp. 6645–6649

[9] A. Graves, N. Jaitly, and A.-R. Mohamed, “Speech recogniton with deep recurrent neural networks,”[A] in Proc. Automatic Speech Recognition and UnderstandingWorkshop (ASRU)[C], Olomouc, Czech Republic, 2013, pp. 273–278

[10] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,”[A] in INTERSPEECH[C], 2014

[11] S. Hochreiter and J. Schmidhuber, “Long short-term memory,”[J] Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997

[12] Felix A. Gers, Nicol N. Schraudolph, and J ürgen Schmidhuber. “Learning precise timing with LSTM recurrent networks,”[J] Journal of Machine Learning Research, vol. 3, pp. 115–143, Mar. 2003.

Speech Recognition Research on the Fusion of DNN and RNN Acoustic Model

Huifeng Zhu, Yong He, Kun Yong, Pei Ding, Jie Hao

Toshiba (China) Co., LTD. Research & Development Center, Beijing 100600 China

Abstract: Speech recognition system based on context-dependent neural network hidden Markov models(CD-NN-HMM) have become the mainstream in commercial field. It has been proved that CD-NN-HMM significantly outperforms the conventional Gaussion mixture model(GMM). Deep neural network and recurrent neural network are the two representative models of CD-NN-HMM architecture. In This paper, a novel method of speech recognition based on the fusion of DNN and RNN in the topological structure and the acoustic score is studied. The experimental results show the acoustic model fusion can effectively improve the speech recognition accuracy.

Key words: DNN; RNN; Speech Recognition; Acoustic Model Fusion