基于循环神经网络的语音识别模型

朱小燕 王 昱 徐 伟

(清华大学智能技术与系统国家重点实验室 北京 100084)

(清华大学计算机科学与技术系 北京 100084)

摘 要 近年来基于隐马尔可夫模型(HMM)的语音识别技术得到很大发展.然而 HMM 模型有着一定的局限性,如何克服 HMM 的一阶假设和独立性假设带来的问题一直是研究讨论的热点.在语音识别中引入神经网络的方法是克服 HMM 局限性的一条途径.该文将循环神经网络应用于汉语语音识别,修改了原网络模型并提出了相应的训练方法.实验结果表明该模型具有良好的连续信号处理性能,与传统的 HMM 模型效果相当.新的训练策略能够在提高训练速度的同时,使得模型分类性能有明显提高.

关键词 语音识别,隐马尔可夫模型(HMM),循环神经网络

中图法分类号: TP391

Speech Recognition Model Based on Recurrent Neural Networks

ZHU Xiao-Yan WANG Yu XU Wei

(*State Key Laboratory of Intelligent System and Technology, Tsinghua University, Beijing* 100084)

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract To overcome some weaknesses of hidden Markov model in speech recognition, HMM/NN hybrid systems had been explored by many researchers in recent years. In the previous HMM/NN hybrid systems, the neural networks adopted are mostly multilayer perceptron (MLP). In our system, recurrent neural networks (RNN) were used to take the place of MLP as the syllable probability estimator. RNN is MLP incorporated with a feedback which can transport the output of some neurons to other neurons or themselves. The incorporation of feedback into a MLP gives the net the ability to efficiently process the context information of time sequence, which is especially useful for speech recognition. In this paper, the architecture of the RNN is modified and corresponding training schema is presented.

Following techniques have been adopted in our system.

1. A network with a single layer has been adopted, while the content of feedback is different from the network used by previous researchers, i.e., the external output is included in the feedback, not just the internal state output.

2. The training algorithm adopted in our system is back-propagation through time (BPTT) algorithm. In the common BPTT algorithm, the initial feedback values are set arbitrarily according to experience. This means that the initial feedback is not specific to the problem we are dealing with. So it should be preferable if the initial feedback values also can be trained. In our training algorithm, this is achieved by adding an additional layer to the unfolded network.

3. To train the network, proper target values must be given. To acquire them, we take use of HMMs which have been trained to recognize the same syllables. The advantage of this method

收稿日期:1999-12-21.本课题得到国家自然科学基金(69982005)、国家重点基础研究发展规划项目(G199803050703)资助.朱小燕,女,1957 年生,博士,副教授,从事模式识别、文字识别、语音识别及神经网络等方面的研究工作.王 昱,男,1975 年生,硕士研究生,从事语音识别方面的研究.徐 伟,男,1974 年生,学士,现在美国 CMU 大学攻读博士学位,从事语音识别和对话系统的研究.

is that it avoids the difficulty and inaccuracy of the hand-set teacher signals and it gives a smooth transition between two adjacent states.

4. In order to make the network learn faster and acquire better generalization ability, a strategy which trains the network by stages has been used. At first, short fragments of speech sequences are given. After small enough error has been achieved on these short pieces, longer fragments are used to learn. Finally, whole sequences are learned.

Experiment results show that the training speed can be accelerated by the method, and the recognition performance is also improved.

Keywords speech recognition, hidden markov model, recurrent neural networks

1 引 言

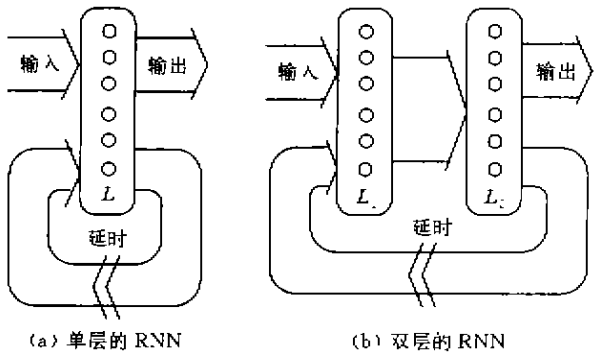
语音对于人来说是一种最自然的交流方式.语音识别技术的研究近年来取得了引人注目的成就.随着隐马尔可夫模型(Hidden Markov Model, HMM)在众多语音系统中的广泛使用,它被普遍认为是目前语音识别领域最成功的模型.但是 HMM 模型也存在着一些自身的局限性.比如声学模型存在量化误差和模型参数假设;标准的最大似然(Maximum Likelihood, ML)训练算法使得声学模型的判别能力降低;一阶假设使得对延迟和协同发音很难模型化;独立性假设则忽略了帧间的相关性.这些局限性使得使用单一的 HMM 模型方法进一步提高性能变得很困难.这样人们开始寻求新的方法.

神经元网络(Neural Networks, NN)是受动物神经系统启发,利用大量简单处理单元互联而构成复杂系统,以便用来解决一些复杂模式识别与行为控制问题.NN 中大量神经元并行分布运算的原理、高效的学习算法以及人的认知系统的模仿能力等都使它极适宜于解决类似语音识别这样的课题.于是人们开始把 NN 和 HMM 的方法结合在一起运用到语音识别中,即 NN/HMM 混合模型[1-5,8]、NN 的引入降低了概率上太强的假设,同时它的训练算法也是可判别的.

目前的 NN/HMM 混合模型系统中,大多数使用的是多层感知器(Multilayer Preceptrons, MLP)网[1-3].国内的 NN/HMM 系统[8,9]也主要以 MLP 为基础,如参考文献[9]就提出了一种反馈的双 MLP 结构.90 年代初开始有人使用了循环神经元网络(Recurrent Neural Networks, RNN)来代替 MLP 进行音子概率估计[5].RNN 是一种既有前馈通路,又有反馈通路的神经元网络.反馈通路的引入,使得网络能够有效的处理时间序列的上下文信息,这对语音识别来说非常重要.本文提出了基于循环神经网络的汉语语音连续识别全反馈模型,采取引入初始层训练的方法提高了系统识别性能和系统稳定性.同时在网络训练算法中提出了采用样本分步训练、教师信号分段添加等的方法,在提高训练速度和效率的同时,使得模型分类性能有明显提高.

2 循环神经元网络

循环神经元网络(RNN)是一种既有前馈通路,又有反馈通路的神经元网络.其中反馈通路可将某一些神经元的输出经过一个或几个时间节拍之后送到其它神经元或自身.反馈通路的引入,使得网络能够有效地处理时间序列的上下文信息,这对语音识别来说是尤其重要的.图 1 中(a),(b)是一些 RNN 的例子[5,7].90 年代初期有人提出利用 RNN 进行语音音素识别[4,5].在此我们针对汉语语音连续识别做了一些研究工作:在原网络模型基础上提出全反馈模型,并提出了初始层训练、教师信号分段添加训练、样本分步训练的学习方法,使得模型分类性能有明显提高.



(a) 单层的 RNN (b) 双层的 RNN

图 1 传统 RNN 的例子

2.1 网络结构

一般 RNN 网络的输出层分为两部分：直接输出部分和反馈部分，如图 1(a). 考虑到连续语音信号的特点，它的每一位输出代表着一个识别单位发生概率的大小，有必要将所有这些信息传递到后续帧信号. 因此，我们提出将所有输出全反馈的模型，如图 2 所示. 实验结果证明此模型优于部分反馈模型.

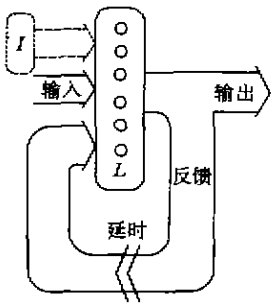


图 2　全反馈的 RNN 的网络结构

从图 2 中也可以看出我们在引入全反馈的同时，在输入层增加了 I 层. 在通常的RNN网络训练算法中，初始时刻的反馈值通常是根据经验人为设定的，并且很多情况下这个值被设为活化函数输出值的中点[7]. 于是初始反馈就与所处理的特定问题无关. 然而事实往往不是如此，尤其是像前后衔接关系密切的语音信号，训练样本与实际样本有出入，或由于数据庞大等关系使得训练样本长度不够的情况下，就会带来较大的误差. 为此，我们引入 I 层. 训练时如图 3 所示，假定层 I 中的每个神经元都连接到一个输出恒为 1 的神经元，那么层 I 的神经元的输出值就由它们和神经元 1 相连权值所决定. 给定一个从层 L 反传过来的误差向量，这些权值就可以和其它权值一样进行修正. 实验结果证明 I 层的引入及相关的训练使得网络性能大幅度提高.

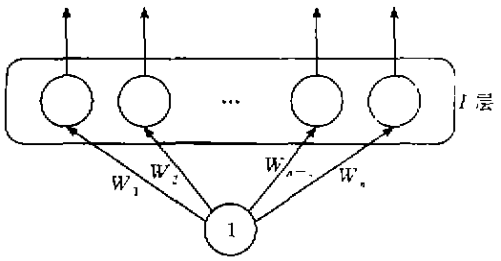
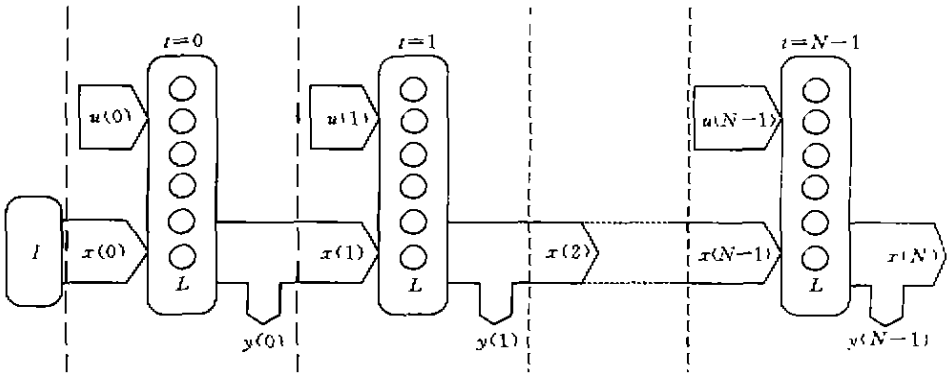


图 3　层 I 的训练

2.2 训练算法

网络的训练采用通过时间的反向传播算法(Back-Propagation Through Time，BPTT)[6]. 该算法的基本思想是把网络按时间展开，并且把展开后的网络看作是一个在各时间步之间有共享权值的大网络. 图 2 的网络展开后的结构如图 4 所示.



其中，$u(t)$为输入向量，$x(t)$为反馈向量(状态向量)、$y(t)$为输出向量

图 4　RNN 网络训练示意图

注意到在展开后的网络中的附加层 I. I 层是用来为层 L 产生初始反馈(状态)信号的. 考虑图 4，对一段长度为 N 的输入输出序列来说，网络训练过程和相应的公式如下：

(1) 把 $x(0)$设为从层 I 输出的初始状态，$u(0)$设为第一个输入. 向前传播计算得到 $y(0)$和 $x(1)$.

(2) 对所有的 $t>0$，把 $x(t)$设为前一个时刻的状态输出，$u(t)$设为当前时刻的输入，向前传播计算得到 $y(t)$和 $x(t+1)$.

$$z(t)=\begin{bmatrix}1\\ u(t)\\ x(t)\end{bmatrix} \tag{1}$$

$$y_i(t+1)=f(W_i z(t)) \tag{2}$$

其中W_i为神经网络权值，$f(x)=\tanh(x)=\frac{2}{1+e^{-2x}}-1$.

以上两步的前向过程即算出循环网络在每个时间节拍的输出. 反复进行直到最后一帧信号输入完毕，$t=N-1$.

下面为反向过程.

(3) 把最后一个时间拍的状态向量的误差置为零,因为目标函数的取值和最后一个状态向量无关.通过和输出的目标值进行比较,得到最后一个时间拍的输出 $y(N-1)$ 误差向量.如同在 MLP 中隐含单元的误差反传一样,把计算出的误差反传到 $x(N-1)$.

$$e_i(N-1)=\begin{cases} f'(W_i z(N-1))\times(y_i(N-1)-o_i(N-1)), & 0\le i<C \\ 0, & \text{其它} \end{cases} \tag{3}$$

其中 C 是待识别的种类,$o_i(t)$ 是目标值.

(4) 对所有的 $0\le t\le N-2$,通过和当前时刻的目标输出进行比较,计算输出向量 $y(t)$ 的误差向量,并且把这个向量加到在 $t+1$ 时刻反向传播过来的误差向量上.然后反向传播计算 $x(t)$ 的误差向量.

$$e_i(t)=\begin{cases} f'(W_i z(t))\times(y_i(t)-o_i(t)+\sum_j w_{ij}e_j(t+1)), & 0\le i<C \\ f'(W_i z(t))\sum_j w_{ij}e_j(t+1), & \text{其它} \end{cases} \tag{4}$$

(5) 在时刻 0,把误差从层 L 反向传播到层 I.

$$e_i^I = f'(w_i^I)\sum_j w_{ij}e_j(0) \tag{5}$$

(6) 对所有时间拍累加计算得出目标函数对权值的梯度,然后更新权值.

$$\Delta w_{ij} = \alpha\sum_{t=0}^{N-1} z_i(t)e_j(t) \tag{6}$$

$$\Delta w_i^I = \alpha e_i^I \tag{7}$$

由式(4)可以看出当训练样本足够长时,误差累积计算项很多.如当语音帧数 600 帧时误差项为 600 次误差的累计,误差总值使得权值产生较大的跳动,甚至产生饱和.解决此问题的方法之一是减小训练系数,而较小的训练系数会影响网络的收敛速度,加大训练时间.因此我们对训练方法做了下面的改进.

首先我们提出阶跃式分步训练的方法.根据语音信号的短时平稳性,对语音训练样本进行跳跃训练,即分阶段进行训练.开始时从样本中抽取 1,3,5,7,…帧进行训练,这样训练的总帧数相应减半;等到训练到一定程度后,再用样本的 0,2,4,6,…帧进行训练;最后才用完整的样本进行训练.这种方法经过对可控二进制计数器的模拟实验(实验 1)和语音识别的训练实验证明可以大大减小训练时间,提高训练质量.

除此之外,传统的 RNN 网络训练是在最小输入单位结束时给以教师信号.但是,作为连续语音信号处理在中间环节产生的误差也比较大.因此,我们采用了中间过程加入教师信号的训练方法,大大加快了训练速度和系统识别性能.

3 RNN/HMM 混合模型的实现

由于在 NN/HMM 混合模型[2-5,8]中,和标准 HMM 模型有很大一部分是完成相同的功能,因此 NN/HMM 可以充分利用标准 HMM 已有的代码.实验系统中 RNN/HMM 混合模型的框架是从 HMM 框架派生而来的,它继承了原来的 HMM 框架的大部分特性.RNN/HMM 和标准 HMM 的主要不同之处是在于标准 HMM 采用混合高斯密度计算状态输出概率 $P(u_t|s_i)$,而 RNN/HMM 是用神经元网络计算这个概率,方法如下:

$$P(u_t|s_i) = \frac{P(s_i|u_t)P(u_t)}{P(s_i)} \tag{8}$$

式中 $P(s_i)$ 可以从训练数据中 s_i 的相对频度统计得到,$P(u_t)$ 跟状态序列无关,对解码过程没有影响,可以忽略,而 $P(s_i|u_t)$ 可以由循环网络来得到.HMM 的转移概率的估计可以直接通过状态序列得到.设 s_i 为状态序列的第 i 个状态.s_{ij} 是状态序列中的状态对 i 与 j,则转移概率为 $P(s_{ij}|s_i)$.首先设定转移概率的初始值,然后计算基于初始值的最佳状态序列,并用该状态序列修改转移概率和神经网络参数.此方法与 Baum-welch 算法相似.整个 RNN/HMM 混合模型识别过程如图 5 所示,其中搜索最佳路径的方法和标准的 HMM 解码过程完全一致.

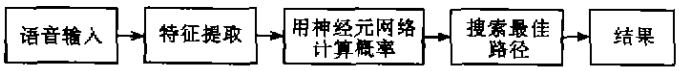


图 5 识别过程

4 实验及结果分析

实验中语音数据取自清华大学智能技术与系统国家重点实验室收集整理的语音数据库 CIDS,用到的数字语音数据包括 60 个人.其中 40 个人的语音作为训练集,其他 20 个人的语音作为测试集.语音数据的采集频率为 11.025KHz,使用的声学特征是

采用 16 维倒谱系数＋16 维差分倒谱系数＋1 维差分能量.

实验 1.　对样本分段训练策略性能的测试.

实验的模拟目标是把 RNN 训练成为一个可控二进制计数器. 为此采用了如图 1(b)的两层网络结构. 输入只有 1 个，作为计数器的控制信号，L_1 层有 5 个单元，L_2 层有 2 个单元. L_2 层的两个输出反馈到 L_1 层作为输入. 该计数器用－1 代表二进制的 0，1 代表二进制的 1，它实现的功能是：当输入为 0 时作减一计数，当输入为 1 时作加一计数.

我们比较了两种训练方式：第一种是一开始就用长度为 40 的输入教师及相应输出(序列是随机产生的，每次所用的序列和以前的都不相同)给网络训练，结果是网络始终不能收敛；第二种采用分段训练的方法，即开始只用长度为 6 的输入输出对给网络训练，等收敛后把长度增加到 8. 这样逐渐增加训练序列的长度直到达到足够长为止，这样训练网络很快就可以对长达几百的输入产生正确的输出了.

这个实验说明对 RNN 的样本分段训练可以大大减小训练时间，提高训练质量. 在后面语音识别的实验中也可以看到这种方法的良好效果.

实验 2.　用 RNN 对数字语音进行识别.

对数字语音信号进行识别实验，以比较各改进训练方法的有效性. 在此我们分别把输出层全反馈的方法 A_{new}、训练初始反馈信号的方法 B_{new}、样本分段训练的方法 C_{new}和教师信号分段添加的方法 D_{new} 四种方法和相应的传统方法 A_{old}，B_{old}，C_{old}，D_{old}组合成不同的训练方案策略，进行了对比实验. 其中各个新老方法具体内容如下：

- 方法 A_{new}：层 L 的所有输出都反馈.
- 方法 A_{old}：层 L 对外界的输出不反馈.
- 方法 B_{new}：对网络的初始反馈进行训练.
- 方法 B_{old}：不对网络的初始反馈进行训练，把它初始反馈设为 0.
- 方法 C_{new}：在输入样本的每一帧都加误差向量，和该样本对应的输出的目标值设为 1，否则为－1.
- 方法 C_{old}：只在输入样本的最后一帧加误差向量，前面各帧的输出误差向量都置为 0.
- 方法 D_{new}：样本分阶段进行训练，开始时从样本中抽取 1，3，5，7，…帧进行训练，等训练到一定程度后用样本的 0，2，4，6，…帧进行训练，最后用完整的样本进行训练.
- 方法 D_{old}：一开始就用完整的样本进行训练.

对比实验是这样进行的. 首先我们在全样本输入的前提下：方案Ⅰ网络结构采用部分反馈网络，以比较网络结构的修正对系统性能的影响；方案Ⅱ采用初始反馈给定值，不对 I 层训练，以比较 I 层引入的效果；方案Ⅲ教师信号只在最后一帧添加，以比较教师信号加入方式的影响；方案Ⅳ则同时在上面三个方法上使用改进的方法，即相对于以上三个含旧方法的方案比较新训练方法的效果. 最后对方案Ⅳ在原改进训练方法的基础上进一步添加上训练样本分段训练的训练策略得到方案Ⅴ，其结果代表了系统的最终性能. 具体的各个方案的实验结果由表 1 给出.

表 1　RNN 进行孤立数字识别，各种方案的比较

	训练策略	训练集的错误率	测试集的错误率
方案Ⅰ	$A_{old}+B_{new}+C_{new}+D_{old}$ 部分反馈，即对外的输出不反馈，且直接使用完整样本训练	1.0%	19.5%
方案Ⅱ	$A_{new}+B_{old}+C_{new}+D_{old}$ 初始反馈给定值不使用 I 层训练，且直接使用完整样本训练	6.8%	18.0%
方案Ⅲ	$A_{new}+B_{new}+C_{old}+D_{old}$ 教师信号只在最后一帧添加，且直接使用完整样本训练	0.0%	21.5%
方案Ⅳ	$A_{new}+B_{new}+C_{new}+D_{old}$ 直接使用完整样本训练	0.25%	5.0%
方案Ⅴ	$A_{new}+B_{new}+C_{new}+D_{new}$ 使用了所有的新方法	0.25%	3.5%

通过方案Ⅰ与方案Ⅳ的比较，可以看出全反馈网络 A_{new}使得训练错误率降低了 75%，测试样张错误率降低了约 75%，大大提高了网络处理连续信号的能力. 方案Ⅱ与方案Ⅳ的比较说明初始反馈训练 B_{new}方法的应用可以显著地降低系统的识别错误率，尤其是对于训练样本，这说明初始反馈信号对网络分类性能的影响是不容忽视的. 通过方案Ⅲ和方案Ⅳ的比较，可以发现方法 C_{new}输入每一帧添加教师信号的方法，虽然使得训练集错误率略有上升，但大大降低了测试集的错误率；最后通过方案Ⅳ和方案Ⅴ的比较，可以看出我们使用方法 D_{new}，即对样本分步训练的方法，不但提高了系统训练时的收敛速度，也有助于性能的提高；所以实验证明用本文提出的 4 种新方法大大提高了系统的性能.

5　结　论

本文提出了一种基于循环神经网络的汉语语音连续识别全反馈模型，并给出了基于 BPTT 训练算

法的多种训练策略.经过实验证明,这种模型可以达到传统的统计模型的识别效果,但计算量大大低于标准 HMM 模型的方法.所采取的初始层训练、样本分步训练、教师信号分段添加等训练策略都能够在提高训练速度和效率的同时,使得模型分类性能有明显提高.本研究证明实现神经网络识别模型的实际应用是可能的.

使用循环神经网络的语音识别方法有进一步发展的潜力.为了提高系统的性能,减少音节重复识别行错误,需要进一步研究网络模型输出的设置.同时,使用更加复杂精确的教师信号,比如利用训练好的 HMM 模型的参数作为教师信号,可能会有更好的效果.

总之,本文表明了基于循环神经网络的汉语语音识别模型有良好的效果,并且有希望进一步提高,成为自然发音的语音识别的新途径.

参 考 文 献

1 Bourlard H, Morgan N. Continuous Speech Recognition: A Hybrid Approach. Norwell, Massachusetts: Kluwer Academic Publishers, 1994

2 Abrash V, Franco H, Sankar A *et al*. Connectionist speaker normalization and adaptation. In: Proc 4th European Conference of Speech Communication and Technology (Eurospeech 95), Madrid, Spain, 1995

3 Cohen M, Rumelhart D, Morgan N *et al*. Combining neural networks and hidden Markov models for continuous speech recognition. In: Proc DARPA Speech and Natural Language Workshop, Harriman, NY, 1992

4 Tebelskis J. Speech using neural networks. Carnegie Mellon University: Technical Report CMU-CS-95-142, 1995

5 Robinson T. An application of recurrent nets to phone probability estimation. IEEE Trans Neural Networks, 1994, 5(3): 298—305

6 Werbos P J. Backpropagation through time: What it does and how to do it. Proceedings of the IEEE, 1990, 78(10): 1550—1560

7 Senior A W. Off-line cursive handwriting recognition using recurrent neural networks[Ph D dissertation]. Cambridge: University of Cambridge, 1994

8 Yu Tie-Cheng, Zhou Jian-Lai, Song Yan-Tao. An overview of speech recognition based on the hybrid NN/HMM approach. In: Proc 5th National Conference of Man-Machine Speech Communication (NCMMSC-98), Harbin, 1998. 18—21(in Chinese)
(俞铁城,周健来,宋岩涛.基于神经网络/隐马尔可夫模型的混合语音识别方法的研究现状.见:第 5 届全国人机语音通讯学术会议论文集,哈尔滨,1998. 18—21)

9 Li Quan-Zai, Chen Dao-Wen. Chinese connected digits speech recognition system based on the hybrid HMM/ANN approach. In: Proc 5th National Conference of Man-Machine Speech Communication (NCMMSC-98), Harbin, 1998. 166—168(in Chinese)
(李全在,陈道文.基于混合 HMM/ANN 方法的汉语连续数字识别系统.见:第 5 届全国人机语音通讯学术会议论文集,哈尔滨,1998. 166—168)