

基于循环神经网络的音素识别研究

黎长江, 胡 燕

(武汉理工大学 计算机科学与技术学院, 湖北 武汉 470070)

摘 要: 基于隐马尔科夫模型(HMM)和循环神经网络(RNN)的 HMM-RNN 混合模型在语音识别中取得了很大的成功. 由于 RNN 的计算过程本身就是上下文相关的, 相邻帧的重合部分增加了整个系统的训练时间. 针对上述问题, 使用连接时序分类(CTC)来代替 HMM 跟 RNN 结合, 并在语音分帧过程中去除相邻帧之间的重合部分, 使用 TIMIT 语音数据集, 进行音素上的识别任务, 并且实验结果表明 CTC-BLSTM 模型在音素上的识别率要高于 HMM-BLSTM 混合模型, CTC-BLSTM 在去除帧重合后能够大幅提高系统的训练效率并且保证识别率大致相同.

关键词: 语音识别; 连接时序分类; 循环神经网络

中图分类号: TP391.42

文献标识码: A

文章编号: 1000-7180(2017)08-0047-05

Research of Phoneme Recognition Based on Recurrent Neural Network

LI Chang-jiang, HU Yan

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Recently HMM-RNN hybrid system has been proved to be successful in speech recognition, since the calculation of RNN is context-dependent, the overlapped part increases the training time. This paper combines CTC with RNN instead of HMM, and remove the overlapped part during framing modeling TIMIT dataset on phone recognition tasks. The experiments show that CTC-BLSTM performs better than HMM-BLSTM on phone recognition, and removing the overlapped part of frames can make system more efficient and ensure the accuracy at a certain degree.

Key words: speech recognition; CTC; RNN

1 引言

在自动语音识别中, 通常使用声学模型和深度神经网络结合的混合方法, 例如 HMM-NN 的模型, 其中神经网络作为识别模型来估计语音发音状态的后验概率. 使用这种方法存在两个问题: 第一, 训练数据必须预先进行对齐处理, 每一段数据都给定一个期望输出; 第二, 由于网络只输出当前时间点标签的分类情况, 从全局来看, 两个标签连续出现的似然度需要另外的模型来给出^[1].

RNN 和 HMM 结合可以用来进行时序分类, 但是同时这种模型继承了 HMM 的缺点^[2]. 首先, HMM 是生成式的, 而神经网络是判别式的, 能够在

分类的问题上做的更好, RNN 相对 HMM 来说提供了更加灵活的输入模型. 其次, 标准 HMM 的内部状态是离散单独的值, 而 RNN 的内部状态是由隐层单元的激活函数确定的, 也就是说 RNN 的信息量随着隐层单元的数量增长而增长. 接着 HMM 需要给输入序列进行分段, 这样才能决定隐藏状态, 但是单元之间的界限非常模糊, 所以给输入序列进行分段比较困难. 最后, HMM 假设下一个状态仅由当前状态所决定, 这意味着有连续轨迹的数据例如手写体很难来用标准的 HMM 来进行建模. 而对于 RNN 来说, 处理这类数据则显得很自然. 因此, 将 HMM 和 RNN 进行结合, HMM 并无法完全利用 RNN 在于长范围序列模型上的优势.

CTC 使得 RNN 可以直接进行序列数据分类的任务,移除了语音数据预对齐的工作,而且 CTC 直接输出所有预测序列的概率,而且不需要后部的额外处理^[3].

使用 RNN 进行语音识别的任务,首先得进行语音预处理,在分帧的过程中,为了使每一帧能够平滑过渡,相邻的帧之间有重合的部分.而由于 RNN 的特殊结构,每个节点的输出和相邻节点的输入相关,所以通过去除帧之间的重合部分,使得在神经网络的通过输入序列的长度减少,大幅降低训练时间.

2 循环神经网络

对于传统的神经网络,我们假定所有的输入都各自相互独立,但是在很多应用场景下,实际情况并不是这样.下一刻所发生的很可能跟这一刻所发生的事件有关,甚至跟一开始到现在发生的所有事件有关,如果需要预测下一刻会发生什么,那么最好也知道之前发生过的事件^[4]. RNN 在计算当前节点的时候,其计算结果依赖于之前节点的计算,图 1 简略描述了 RNN 的结构

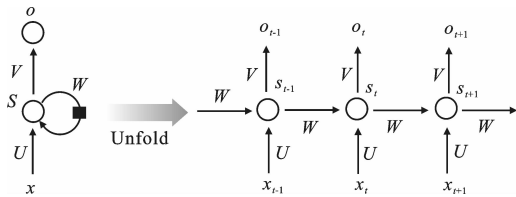


图 1 RNN 模型

如图 1 所示,RNN 的隐层自连接,展开之后得到右图所示的模型。其中 U, V, W 分别为输入层,输出层,隐层的链接权值, x_t 为当前节点的输入,通过它和上一个节点的隐层输出 s_{t-1} 可以计算得到当前节点的隐层输出 s_t ,通过 s_t 则可以计算得到该节点的输出 O_t , O_t 的计算过程大致如下:

$$s_t = \tanh(x_t U + s_{t-1} W + b_s) \tag{1}$$

$$o_t = s_t V \tag{2}$$

在计算第一个节点隐层输出的时候,使用元素全为 0 的向量来初始化 s_0 ,如果是多分类分体,则在计算出 o_t 后增加一步 softmax 的计算:

$$y_t = \text{softmax}(o_t + b_y) \tag{3}$$

2.1 双向循环神经网络(BRNN)

对于序列标注来说,一个节点的标注可能不仅跟它的过去有关系,而且跟它的将来也有关系,好比阅读,一句话或者一个词所要传达的意思是有上下文来决定的,序列标注也是一样^[5].

标准的 RNN(LSTM)以时序来处理序列数据从而忽略了下文,BRNN 的概念在 1997 年被提出来,给除了一个比较合理的解决方案,其中 BRNN 的结构如图 2 所示.

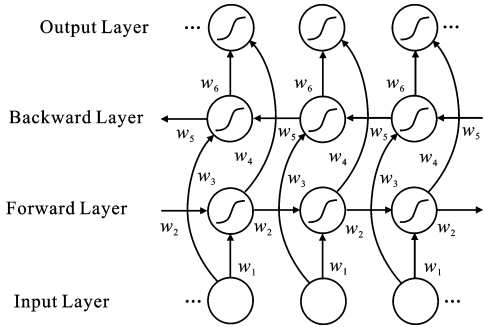


图 2 BRNN 模型

BRNN 的基本思想是让训练序列分别通过一层正向的隐层和一层逆向的隐层,这两个分开的隐层都连接到输出层,通过正向计算得到的 S_t 和反向计算得到的 S_t 计算输出.

2.2 长短时记忆(LSTM)

RNN 认为当前节点的输出跟过去所有的节点有关,但实际情况只是跟过去连续的几个节点有关,而且在使用 BPTT 算法训练 RNN 的时候,会出现梯度消失的问题,所以实验中采用另外一种结构的 RNN,LSTM 来避免这种问题^[6]. 图 3 为 LSTM 隐层单元的结构.

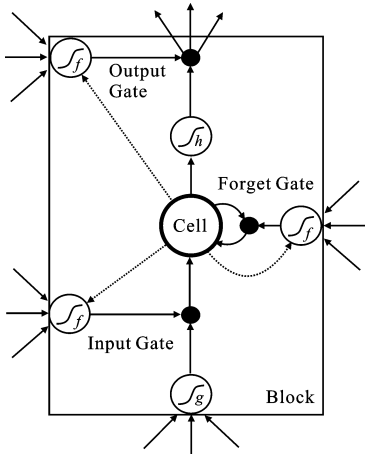


图 3 LSTM 神经元

如图 3 所示,隐层引进了“门”的概念.激活函数 f, g, h 通常是 sigmoid 或者 tanh 函数,虚线表示让门也去接受中间记忆细胞的输出,有的 LSTM 会这样做. s_t 的计算流程大致如下:

$$i = \sigma(x_t U^i + s_{t-1} W^i + b_i) \tag{4}$$

$$f = \sigma(x_t U^f + s_{t-1} W^f + b_f) \tag{5}$$

$$o = \sigma(x_t U^o + s_{t-1} W^o + b_o) \quad (6)$$

$$g = \tanh(x_t U^g + s_{t-1} W^g + b_g) \quad (7)$$

$$c_t = c_{t-1} \cdot f + g \cdot i + b_c \quad (8)$$

$$s_t = \tanh(c_t) \cdot o \quad (9)$$

LSTM为RNN的一种变种,改变了隐层单元的结构和 s_t 的计算过程,从上面的公式可知,sigmoid通过限制隐层单元每个“门”的输出从而限制整个隐层单元是否输出,也就是合适记忆,何时忘记。

3 基于BLSTM循环神经网络语音识别模型

RNN擅长处理序列数据,但是对于语音识别的任务来说,使用RNN进行此类任务需要预先对语音数据进行分段,而且需要对语音的输出标签进行预处理,前者好解决,但是后者工作量则巨大.语音识别声学模型的训练属于监督学习,需要知道每一帧对应的标签才能进行有效的训练,使用连接时序分类(CTC)则不用考虑上面的问题,只要一个输入序列和输出序列就能进行有效的训练,CTC直接输出预测序列的概率,不需要外部的后处理. CTC的其他不同之处在于引进了一个blank(—)标签,表示空白,去掉输出中连续重复的元素和空白,就能得到标签,输出和标签满足以下的等价关系:

$$F(a - a b -) = F(- a a - a b b) = aab$$

CTC接在RNN的softmax层之后,RNN每个节点的softmax层都能得到一个向量,表示各个标签在此刻的概率,根据每个节点的各个标签的条件概率,可以计算出每一条可能路径的条件概率,假设输入 X 的长度为 T ,标签长度为 U ,那么可能路径的数量有 U 的 T 次方个,在CTC中需要将标签首尾和标签之间插入blank也就是这个数量 L 的 $2U+1$ 次方个,CTC在计算条件概率时引进了类似HMM中的前后向算法,简化了计算流程。

图4描述了基于CTC-BLSTM的语音识别模型,将语音预处理阶段得到的MFCC参数,作为系统的输入,最后经过Softmax层得到输出。

在训练阶段将Softmax层的输出作为CTC层的输入,通过CTC算法计算出损失值,然后通过BPTT算法,调整网络的参数.在测试阶段则将通过集束搜索方法,对Softmax层的输出进行解码,得到对应的输出序列^[7]。

3.1 前后向算法

输入 X 的长度为 T ,将长度为 U 的标签首位以及标签直接插入blank,得到长度为 $U' = 2U + 1$ 的

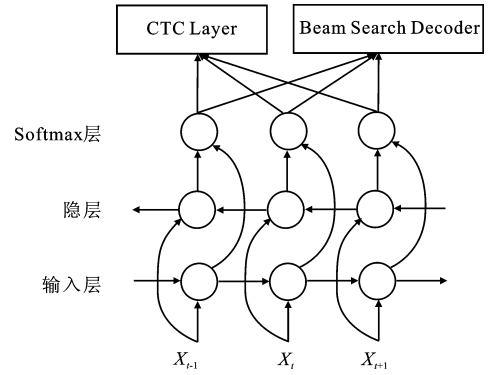


图4 CTC-BLSTM 语音识别模型

标签集合 l ,前向变量为 $\alpha(t, u)$,表示 t 时刻在节点 u 的前向概率值, $\beta(t, u)$ 为后向变量,表示 t 时刻在节点 u 的概率值,其中 $u \in 2U + 1$, y_k^t 表示时刻 t 输出 k 的概率值.前后向变量初始化及递推关系如下

$$\alpha(1, 1) = y_b^1 \quad (10)$$

$$\alpha(1, 2) = y_{l_1}^1 \quad (11)$$

$$\alpha(1, u) = 0, \forall u > 2 \quad (12)$$

考虑到时间的限制,有些节点并不能到达最后终止的节点,所以有

$$\alpha(t, u) = 0, \forall u < U' - (2T - t) - 1 \quad (13)$$

同样增加一个边界变量

$$\alpha(t, 0) = 0 \quad (14)$$

前向变量的递推关系如下

$$\alpha(t, u) = y_{i_u}^t \sum_{i=f < u} \alpha(t-1, i) \quad (15)$$

$$f(u) = \begin{cases} u-1 & \text{if } l_u = \text{blank or } l_u = l_{u-2} \\ u-2 & \text{otherwise} \end{cases} \quad (16)$$

$$\beta(T, U') = 1 \quad (17)$$

$$\beta(T, U' - 1) = 1 \quad (18)$$

$$\beta(T, U') = 0 \quad (19)$$

同样增加一个边界变量

$$\beta(T, U' + 1) = 0 \quad (20)$$

后向变量的递推关系如下

$$\beta(t, u) = \sum_{i=g < u} \beta(t+1, i) y_{i_u}^{t+1} \quad (21)$$

$$g(u) = \begin{cases} u+1 & \text{if } l_u = \text{blank or } l_u = l_{u+2} \\ u+2 & \text{otherwise} \end{cases} \quad (22)$$

计算完前后向变量后,定义 $z = l, B(z, k) = \{u; z_u = k\}$, a^t 为RNN在时间 t 的softmax层的输入,在任何时间节点 t ,求出所有节点的前后向变量的乘积的和,然后对数取反,可以得到该时间节点的误差值 $L(x, z)$,接下来则可以利用BPTT算法沿着神经网络反方向传播误差^[8],

$$p(z | x) = \sum_{u=1}^{|z|} \alpha(t, u) \beta(t, u) \quad (23)$$

$$L(x,z)=-\ln \sum_{u=1}^{|z|} \alpha(t,u) \beta(t,u) \tag{24}$$

$$\frac{\partial L(x,z)}{\partial a_k^t} = y_k' - \frac{1}{p(z|x)} \sum_{u \in B(z,k)} \alpha(t,u) \beta(t,u) \tag{25}$$

3.2 解码

当完成 CTC 的训练之后,需要有对应的算法对其进行解码. 由于 RNN 最后的 Softmax 层输出了每一个时间节点的各个音素的概率值,其中一种比较简单的方法是选取每一个时间节点最大概率输出:

$$\operatorname{argmax}_y \Pr(y|x) \approx B(\operatorname{argmax}_\alpha \Pr(\alpha|x)) \tag{26}$$

但是这种解码方式的精度并不高,比较好的方法是采用相关搜索算法来找出最优音素序列^[9]. 算法采用集束搜索,定义

$$\Pr(y,t) = \Pr^+(y,t) + \Pr^-(y,t) \tag{27}$$

$$\Pr(k,y,t) = \Pr(k,t|x) \Pr(k|y) \tag{28}$$
$$\begin{cases} \Pr^-(y,t-1) & \text{if } y^e = k \\ \Pr(y,t-1) & \text{otherwise} \end{cases}$$

式中, $\Pr^+(y,t)$, $\Pr^-(y,t)$ 分别代表输出序列 y 的非空白概率和空白概率, $\Pr(y,t)$ 则为总概率, $\Pr(k,y,t)$ 表示在时间 t 使用标签 k 来扩展 y 的扩展概率, $\Pr(k,t|x)$ 为时间 t 输出 k 的概率值, $\Pr(k|y)$ 为从 y 到 $y+k$ 的过渡概率, y^e 是 y 中的最后一个标签, $\hat{y}$ 是 y 移除最后一个标签之后的序列,集束搜索的过程如下:

CTC 集束搜索

```
Initialize:  $B = \{\varphi\}$ ,  $\Pr^-(\varphi,0) = 1$ 
for  $t$  from 1 to  $T$  do
     $A =$  the  $W$  most probable sequence in  $B$ 
     $B = \{\}$ 
    for  $y$  in  $A$  do
        if  $y \neq \varphi$  then
             $\Pr^+(y,t) = \Pr^+(y,t-1) \Pr(y^e, t|x)$ 
            if  $y$  in  $A$  then
                 $\Pr^+(y,t) = \Pr^+(y,t-1) + \Pr(y^e, y,t)$ 
             $\Pr^-(y,t) = \Pr^-(y,t-1) \Pr(-,t|x)$ 
            Add  $y$  to  $B$ 
    for  $k$  from 1 to  $K$ 
         $\Pr^-(y+k,t) = 0$ 
         $\Pr^+(y+k,t) = \Pr(k,y,t)$ 
    Add  $y+k$  to  $B$ 
return  $\max_{y \in B} \Pr(y|T) / |y|$ 
```

3.3 误差度量

在网络训练完成后,需要对整个系统进行评估,采用编辑距离 ED 作为期望输出 y 和实际输出 y' 的基本误差度量:

$$\text{LER}(y,y') = \sum_{y \in S} \frac{\text{ED}(y,y')}{|S|} \tag{29}$$

4 实验

实验采用 TIMIT 语音数据库进行建模, TIMIT 语音数据集包含了 6 000 多个词汇,朗读者来自美国 8 个不同区域共 630 人,每人朗读 10 句话,一共 6 300 条数据. 系统采用音素(音标)作为基本识别单元, TIMIT 数据集中包含了 61 个音素,事实上英语基本音素只有 48 个. 由于某些音素发音相近,甚至是不发音,对于计算机来说可以弱化这些相近音素所存在的差异,将发音相近的因素看为同一个音素,所以实验将 61 个音素压缩简化为 38 个音素(不含 sil),对应如表 1 所示.

表 1 音素简化表

aa	aa, ao
ah	ah, ax, ax-h
er	er, axr
hh	hh, hv
ih	ih, ix
l	l, el
m	m, em
n	n, en, nx
ng	ng, eng
sh	sh, zh
uw	uw, ux
sil (blank)	pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi, q

TIMIT 数据集的采样率为 16 000 HZ,实验选取 MFCC 作为语音的特征参数,其中语音预加重参数为 0.97,每一帧为 512 个采样点,帧移为 341,每一帧加上汉明窗,梅尔滤波器组频率为 40 Hz 到 8 000 HZ,滤波器的数量为 40 个, MFCC 参数为 13 阶,然后计算其一阶差分和二阶差分,得到共 39 阶的特征参数.

4.1 CTC-BLSTM 和 HMM-BLSTM 的对照试验

实验使用开源的 kaldi 工具箱进行音素识别任务,神经网络使用 BLSTM,隐层数量为 1 层,隐层单元数量为 150,神经网络的输入单元数量为 39,输出单元数量为 39,学习率为 0.008. 分别在上下文相关的 HMM, HMM-BLSTM 和 CTC-BLSTM 下进行实验,每个模型运行 5 次.

由表 2 格的实验结果可以知道, HMM-BLSTM

混合模型表现是要优于 HMM 的,而使用 CTC 进行训练,效果优于混合模型.

表 2 实验结果

模型	错误率/%
HMM	38.5
HMM-BLSTM	34.5±0.07
CTC-BLSTM	32.7±0.06

4.2 去除相邻帧重合的对照试验

实验采用 TensorFlow 作为神经网络的框架,并且使用 CUDA 技术对神经网络的运算过程进行加速,每个模型运行 5 次.

神经网络使用 BLSTM,隐层数量为 1 层,隐层单元数量为 150,神经网络的输入单元数量为 39,输出单元数量为 39,学习率为 0.008.

对照数据同样选取 36 阶 MFCC 作为特征参数,这部分数据每一帧依然为 512 帧,相邻帧无重合部分,实验结果从预处理的时间,系统训练时间,识别错误率来对比两部分数据的性能.

表 3 格为该次实验最终的结果,由于 1 号数据比 2 号数据多了预加重的操作,而且在分帧的阶段,1 号数据中相邻帧有 1/3 是重合的,因此同一段语音最终得到的特征参数向量的长度要比 2 号数据中的要长.从表 3 格得到的结果可以看出,无论是提取特征参数还是训练的过程,1 号数据所花的时间比 2 号数据所花的时间要长大约 50%,但是最后得到的错误率仅比后者要少 2.1%.

表 3 实验结果

语音信号 预处理	参数提取 时间 / 分	训练时间 / 分	错误率 /%
# 1 预加重 分帧 1/3 帧 移加汉明窗	41	98	32.7±0.06
# 2 仅分帧 无帧重合	26	65	34.8±0.11

5 结束语

从实验结果看来,RNN 可以直接用来进行序列数据的建模,搭配用于序列建模的连接时序分类算法,不需要对语音预先对齐,而且效果也要优于要优于 HMM 和 LSTM 的混合模型.并且在语音预处理方面,可以适当减少或者直接省去帧与帧之间的重叠部分,从而减少整个系统的训练时间.

未来的研究将从以下几个方面进行展开:(1)由于神经网络对于输入的灵活性,使用多种语音特征

参数作为输入,提高识别率(2)使用更复杂的网络模型,以及引入语言模型,提高识别率.

参考文献:

[1] Graves A. Neural networks[M]. Berlin Heidelberg: Springer, 2012: 15-35.

[2] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international conference on Machine learning. Carnegie Mellon University, Pittsburgh, USA, ACM, 2006: 369-376.

[3] Sak H, Senior A, Rao K, et al. Fast and accurate recurrent neural network acoustic models for speech recognition[J]. Computer Science,2015:1507. 06947.

[4] 涂俊辉, 续晋华. 基于 HTK 的连续语音识别系统及其在 TIMIT 上的实验[J]. 现代计算机(下半月版), 2009,24(11): 29-33.

[5] Maas A L, Xie Z, Jurafsky D, et al. Lexicon-free conversational speech recognition with neural networks [C]//NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA,2015:345-354.

[6] 周盼. 基于深层神经网络的语音识别声学建模研究 [D]. 合肥:中国科学技术大学, 2014.

[7] Miao Y, Gowayyed M, Metze F. EESSEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Thomas, Virgin Islands, IEEE, 2015: 167-174.

[8] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. Vancouver, British Columbia, Canada, IEEE, 2013: 6645-6649.

[9] Chan W, Jaitly N, Le Q V, et al. Listen, attend and spell[J]. Computer Science,2015:1508. 01211.

作者简介:

黎长江 男,(1991-),硕士研究生.研究方向为语音识别.
E-mail:jonathan.l@qq.com.

胡燕 女,(1965-),教授,硕士生导师.研究方向为智能信息检索.