一种多尺度前向注意力模型的语音识别方法

唐海桃,薛嘉宾,韩纪庆
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 注意力模型是当前语音识别中的主流模型,然而其存在一个缺点,即当前时刻的注意力模型可能产生异常得分.为此,本文首先提出前向注意力模型,其采用上一时刻正常注意力得分平滑当前时刻异常得分.接着通过对上一时刻的注意力得分添加约束因子来对前向注意力模型进行优化,达到自适应平滑的目的.最后,在优化模型基础上提出多尺度前向注意力模型,其通过引入多尺度模型来对不同等级的语音基元进行建模,进而将所得到的不同等级目标向量进行融合,以达到解决注意力得分异常值的目的.采用 SwitchBoard 作为训练集,Hub5'00 作为测试集进行实验,相比于基线系统,多尺度前向注意力模型的词错误率(Word Error Rate,WER)相对降低 14.28%.

关键词: 前向注意力机制;自适应平滑;多尺度;语音识别
中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2020)07-1255-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.07.002

A Method of Multi-Scale Forward Attention Model for Speech Recognition

TANG Hai-tao, XUE Jia-bin, HAN Ji-qing
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Attention-based model is a popular model in speech recognition, however it has a disadvantage that the attention-based model may produce abnormal scores. To solve this problem, this paper first proposes a forward attention model, which adopts normal attention score at the previous moment to smooth the abnormal score at the current moment. Then, the model is optimized to add constraint factors to the attention score at the previous moment to achieve the purpose of adaptive smoothing of the above abnormal scores. Then, a multi-scale forward attention model is proposed on the above model. This model introduces a multi-scale method to model the speech primitives of different levels, and then fuses the target vectors of different levels to solve the outliers of attention score. In the experiment, SwitchBoard is adopted as the training set and Hub5'00 as the test set. Compared with the baseline system, the Word Error Rate (WER) of the proposed system decreased by 14.28% relatively.

Key words: forward attention mechanism; adaptive smoothing; multi-scale; speech recognition

1 引言

近年来,深度神经网络在图像处理[1,2]、语音识别[3,4]、自然语言处理[5,6]等领域都取得了重要的成就.在语音识别中,基于注意力机制的编解码模型[7,8],由于结构简单、模型训练方便、识别效果好而受到广泛青睐.因此,本文主要针对基于注意力机制的编解码语音识别模型进行研究.

虽然传统注意力模型是当前语音识别中较为主流的模型,但其仍然存在语音帧对齐效果较差的问题.为了解决该问题,目前学术界提出了很多改进方法.例如,Zeyer 等采用长短时记忆神经网络(Long Short Term Memory,LSTM)作为语言模型添加到注意力机制中,结果表明语言模型对识别结果影响不大[9].Merboldt 等从注意力机制本身出发,通过引入卷积窗约束,将之前的全局注意力机制向局部进行转换[10].Bahar 等对注意力

收稿日期:2019-06-06;修回日期:2020-02-29;责任编辑:孙瑶
基金项目:国家重点研发项目(No.2017YFB1002102)

模型编码部分改进,采用更复杂的 2DLSTM(Two-Dimensional Long Short Term Memory)作为编码模型,以挖掘语音帧中更加潜在的信息[11].除注意力模型外,连接时序分类模型(Connectionist Temporal Classification,CTC)也是语音识别较流行的模型.Zweig 等在 CTC 中采用不同的语言模型进行比较,其结果不如注意力模型[12].虽然文献[9~11]改进了注意力模型,但都未从注意力模型产生的注意力得分可能存在异常的角度来解决帧对齐的问题.

对于注意力得分中存在异常值的问题,它会导致注意力模型关注较为离散的语音帧,同时导致相邻时刻之间被关注的语音帧位置偏差较大[13~15],典型示例如图 1 所示.音素 O 与第 2 帧相对应,那么音素 W 所对应的语音帧应该在第 2 帧附近,而不是第 9 帧.同理,音素 Y 所对应的语音帧也应该在第 7 帧的附近,而不是第 2 帧.

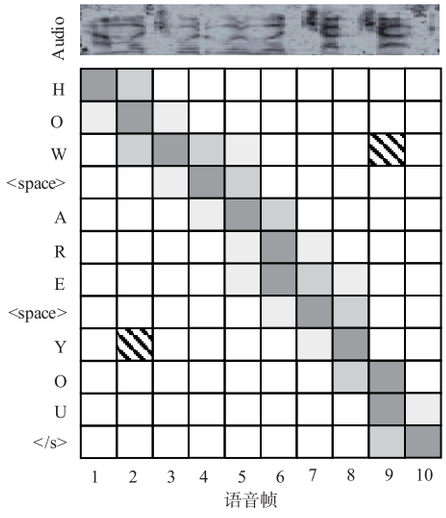

图1 传统注意力机制的异常注意力得分问题示例

为了解决这种问题,本文利用前向算法,提出一种前向的注意力模型,其在计算当前时刻注意力得分归一化后,考虑前后时刻关系,利用上一时刻的注意力得分对当前时刻存在异常得分的语音帧进行平滑.同时考虑到上一时刻每一帧的影响程度不同,又对其进一步优化,利用神经网络计算出上一时刻不同语音帧的约束因子,并根据这些因子与上一时刻的注意力得分来自适应平滑当前时刻的异常点,从而能更好地保证上一时刻重要的语音帧所对应的注意力得分在当前时刻能得到更好的学习.

虽然前向注意力模型在一定程度上缓解了语音帧对齐的问题,但是该方法利用上一时刻的注意力得分对当前时刻进行平滑,只能消除部分异常点.于是,本文进一步深入分析注意力模型的计算过程,发现其只采用单头注意力模型进行建模,导致模型表达能力不够.虽然近年出现的多头(multi-head)注意力模型[16]在一定程度上缓解了该问题,但其只采用单一尺寸的卷积滤波器,来得到固定时长的语音变化模式,其对应输出音素所构成语音基元模型固定不变.

受文献[17]启发,语音中包含不同等级的语音基元,单一尺寸的卷积滤波器不能深入挖掘这些信息.因此,本文在前向注意力模型的基础上采用多头注意力模型,并针对每个头采用不同尺寸的卷积滤波器,提出多尺度前向注意力模型.该模型采用不同尺寸的卷积滤波器,来获取不同时长的语音模式,进而对不同等级的语音基元建模.最终将这些不同等级的模型采用神经网络进行融合,得到包含不同信息的语音基元模型,相比于单一尺寸卷积滤波器的产生模型,能计算出更好的注意力得分.

2 传统注意力模型的语音识别

在语音识别中,基于注意力机制的编解码模型能够解决输入和输出序列长度不相等的问题[18],其将 I 帧的输入语音特征序列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_I)$ 转化为输出文本序列 $\boldsymbol{O}=(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_j,\cdots,\boldsymbol{o}_J)$.这里,$\boldsymbol{x}_i$ 为第 i 帧特征向量;$\boldsymbol{o}_j$ 为当前 j 时刻解码器输出;每个 $\boldsymbol{o}_j$ 可能对应一个或者多个 $\boldsymbol{x}_i$.

首先在编码器部分,输入的语音特征序列 $\boldsymbol{X}$ 能够通过编码器生成更适合注意力机制处理的特征序列 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_I)$:

$$\boldsymbol{H}=\text{Encoder}(\boldsymbol{X}) \tag{1}$$

其中,Encoder(·)是编码器部分,它通常由双向长短时记忆神经网络(Bidirectional Long Short-Term Memory,BLSTM)组成.$\boldsymbol{H}\in\mathbb{R}^{D\times I}$ 为编码器输出特征序列,D 为 BLSTM 神经元个数,即输出特征序列维度,$\boldsymbol{h}_i$ 为第 i 帧编码器输出特征序列.

然后在注意力部分,利用编码器、上一时刻解码器和注意力部分的信息计算当前时刻的注意力得分 $\boldsymbol{\alpha}_j$:

$$\boldsymbol{\alpha}_j=\text{Attend}(\boldsymbol{H},\boldsymbol{q}_{j-1},\boldsymbol{\alpha}_{j-1}) \tag{2}$$

其中,Attend(·)为注意力部分,$\boldsymbol{q}_{j-1}$ 为上一时刻解码器的状态,$\boldsymbol{\alpha}_{j-1}$ 为上一时刻注意力得分.紧接着,利用 $\boldsymbol{\alpha}_j$ 整合 $\boldsymbol{H}$ 得到解码器新的输入,即目标向量 $\boldsymbol{c}_j$:

$$\boldsymbol{c}_j=\sum_{i=1}^{I}\alpha_{i,j}\boldsymbol{h}_i \tag{3}$$

最后在解码器部分,利用 $\boldsymbol{c}_j$、$\boldsymbol{q}_{j-1}$ 和 $\boldsymbol{o}_{j-1}$ 通过解码器得到 $\boldsymbol{o}_j$:

$$\boldsymbol{o}_j=\text{Decoder}(\boldsymbol{o}_{j-1},\boldsymbol{q}_{j-1},\boldsymbol{c}_j) \tag{4}$$

其中,Decoder(·)为解码器部分,它通常由 LSTM 组成.$\boldsymbol{o}_j\in\mathbb{R}^s$ 为当前时刻的输出序列,s 为音素个数.

本文在注意力部分采用局部注意力机制计算当前

时刻注意力得分 $\boldsymbol{\alpha}_j$,其具体计算如下[19]:

首先对上一时刻注意力得分 $\boldsymbol{\alpha}_{j-1}$进行卷积操作:

$$\boldsymbol{f}_j = \boldsymbol{F} * \boldsymbol{\alpha}_{j-1} \tag{5}$$

其中,$\boldsymbol{F}$ 是大小为 k 的卷积窗,$*$ 为卷积.

然后利用卷积后的结果 $\boldsymbol{f}_j$ 和 $\boldsymbol{q}_{j-1}$、$\boldsymbol{h}_i$ 计算当前 j 时刻第 i 帧对应的值 $e_{i,j}$:

$$e_{i,j} = \boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}\boldsymbol{q}_{j-1} + \boldsymbol{V}\boldsymbol{h}_i + \boldsymbol{U}f_{i,j} + b) \tag{6}$$

其中,$\boldsymbol{W}$,$\boldsymbol{V}$,$\boldsymbol{U}$ 分别为解码器、特征序列、注意力得分卷积后对应的权值,b 为偏置.

最后对式(4)的结果利用 Softmax 作归一化处理:

$$\begin{aligned}\alpha_{i,j} &= \mathrm{Softmax}(e_{i,j}) \\ &= \exp(e_{i,j}) \Big/ \sum_{i=1}^{L}\exp(e_{i,j})\end{aligned} \tag{7}$$

其中,$\alpha_{i,j}$为当前 j 时刻第 i 帧的注意力得分值,$0 \leqslant \alpha_{i,j} \leqslant 1$, $\sum_{i=1}^{L}\alpha_{i,j} = 1$.

通过对局部注意力机制进行观察发现,在注意力得分的计算过程中,没有采用任何约束,注意力得分会出现异常值,使相邻时刻之间被关注的语音帧位置偏差较大,从而导致语音帧对齐较差.

3　基于前向注意力模型的语音识别模型

为解决这种异常注意力得分的问题,本文提出前向注意力机制. 它在式(5)之后引入前向算法,利用上一时刻正常的注意力得分平滑当前时刻新的注意力得分,记为 $\hat{\alpha}_{i,j}$. 同时为简化计算,只考虑由前一时刻被关注的语音帧及其附近帧之间的关系,从而提高平滑的效率. 具体公式如下:

$$\hat{\alpha}'_{i,j} = \Big(\sum_{k=0}^{l-1}\hat{\alpha}_{i-k,j-1}\Big) \times \alpha_{i,j} \tag{8}$$

其中,$\hat{\alpha}_{i,j-1}$为上一时刻前向注意力模型计算的注意力得分.

最后利用式(7)的 Softmax 函数对式(8)之后的结果进行归一化处理,得到前向注意力机制的注意力得分 $\hat{\alpha}_{i,j}$.

通过这种方法能利用上一时刻正常注意力得分来对当前时刻的异常值进行平滑,以消除这些异常点,且还能保证前后时刻注意力得分对应语音帧之间的连续性.

但是在前向注意力模型中,$\hat{\alpha}_{i,j-1}$在前 l 个语音帧的影响程度并非一致,且上一时刻被关注的语音帧在当前时刻不可能永远相同,需要对上一时刻前 l 个语音帧添加新的约束,以提高平滑异常值的效果. 在此基础上,又进一步提出优化前向注意力机制,通过采用神经网络(Neural Networks,NN)产生的约束因子 $\boldsymbol{u}_j$ 动态控制上一时刻不同语音帧对应的注意力得分对当前时刻的影响:

$$\boldsymbol{u}_j = \mathrm{NN}(\boldsymbol{q}_{j-1}, \boldsymbol{c}_{j-1}, \boldsymbol{o}_{j-1}) \tag{9}$$

其中,$\boldsymbol{u}_j \in \mathbb{R}^l$为当前时刻通过 NN 利用 $\boldsymbol{q}_{j-1}$,$\boldsymbol{c}_{j-1}$,$\boldsymbol{o}_{j-1}$得到的结果. NN(·)实际上是含有一个隐含层,输出激活函数为 Sigmoid 的神经网络模型.

采用式(9)的约束因子能对上一时刻注意力得分添加新的约束. 这样原本注意力得分较高的语音帧重要程度可能降低,反之,得分较低的语音帧此刻可能就会比较的重要. 通过这种动态调节上一时刻注意力得分的重要程度,达到更好地平滑当前时刻的异常注意力得分的目的. 最终,新的平滑计算公式如下:

$$\hat{\alpha}''_{i,j} = \Big(\sum_{k=0}^{l-1}u_{k,j}\hat{\alpha}_{i-k,j-1}\Big) \times \alpha_{i,j} \tag{10}$$

这里,与式(8)不同的是上一时刻前 l 帧注意力得分通过乘上不同的权值系数达到约束目的.

最后,利用式(5)的 Softmax 函数对式(10)的结果进行归一化,得到当前时刻新的注意力得分 $\hat{\alpha}_{i,j}$. 这样能够更好地保证上一时刻重要的语音帧注意力得分在当前时刻得到更好的学习,实现对当前时刻的异常注意力得分自适应平滑,提升模型的帧对齐效果.

4　基于多尺度前向注意力模型的语音识别模型

虽然第 3 节中的前向注意力模型在一定程度上缓解了注意力得分异常的问题,但只能消除部分的异常值. 对第 2 节中的注意力机制部分进行深入分析发现,其只采用单头注意力模型进行建模,导致模型表达能力不强,这种问题也会导致当前时刻存在异常值.

本文提出多尺度前向注意力模型,该模型利用集成思想,将前向注意力模型采用不同尺寸的卷积滤波器来得到多头模型,并对每个头计算各自注意力得分. 与传统的多头注意力不同,其采用不同大小的卷积滤波器来获取不同时长语音的变化模式,进而能针对不同等级语音基元进行建模,相比传统多头模型采用单一尺度滤波器对固定等级语音基元进行建模,其能够挖掘更加深层和丰富的语音信息. 具体如图 2 所示.

由图 2 可知,在多尺度模型中,对第 3 节中的前向注意力机制得分 $\hat{\alpha}_{j-1}$采用不同尺度的卷积滤波器 F_m 进行卷积计算:

$$\boldsymbol{f}_j = \boldsymbol{F}_m * \hat{\boldsymbol{\alpha}}_{j-1} \quad m = 1, \cdots, M \tag{11}$$

这里,与式(5)不同的是卷积部分采用 M 个不同尺寸的滤波器. 尺寸较小的模型代表着音素一级的模型,正常大小的模型代表着音节一级的模型,而较大的代表着词一级的模型. 由于不同尺寸的卷积滤波器对应着不同大小的滑动窗,在沿着语音帧滑动时尽可能保证每次包含的语音帧对齐的结果能构成一个完整的音

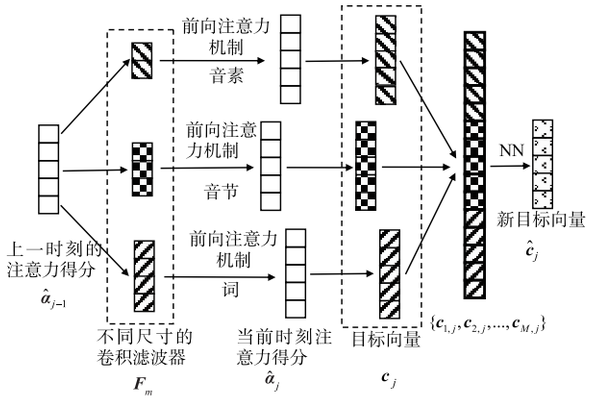

图2 多尺度前向注意力机制的语音识别模型

素,从而防止将同一个音素分割.

在多尺度模型的每个注意力机制中,将式(11)的结果通过式(10)计算出前向注意力得分,进而通过式(3)得到代表不同等级语音基元模型的目标向量 $\boldsymbol{c}_{m,j}$. 考虑到有 M 个不同的目标向量,将其拼接并采用含有一个隐含层的 NN 进行整合,从而得到对语音模型表达更加丰富的目标向量 $\hat{\boldsymbol{c}}_j$:

$$\hat{\boldsymbol{c}}_j = \mathrm{NN}(\{\boldsymbol{c}_{1,j}, \boldsymbol{c}_{2,j}, \cdots, \boldsymbol{c}_{m,j}, \cdots, \boldsymbol{c}_{M,j}\}) \tag{12}$$

采用这种方法,不仅能通过对不同等级语音基元进行建模来得到更好的注意力得分,还能利用上一时刻正常的注意力得分平滑异常值,最终达到较好地消除异常注意力得分的目的,缓解语音帧对不齐的问题.

5 实验结果及分析

5.1 实验数据库

本文采用 Switchboard-II(LDC97S62)数据集[20]进行实验对比与评估. 该语料库共计 2435 段对话,包含 241 位女性与 303 位男性,总时长大约 300h. 对于测试集部分,采用 Hub5'00 数据集,其由 Switch-board("SWB")和 CallHome("CH")数据集组成. 其中,总长为 10h 的 SWB 数据集相对较容易识别,总长为 30min 的 CH 相对较难识别[9~12]. 实验分别用 SWB、CH 和它们的混合数据集作为测试部分. 采用词错误率(Word Error Rate,WER)作为语音帧对齐效果评价指标.

5.2 实验参数设置

所采用的注意力机制模型为含有编码器和解码器的 LAS(Listen Attention and Spell)结构. 在编码器部分,采用 6 层的 BLSTM,每一层的神经元为 320 个. 在保证不对识别结果造成太大影响的情况下,为解决输入语音特征的长时序问题,加速网络的训练,采用每一层跳帧的训练策略,即 1-2-2-1-1-1. 解码器部分为一层含有 300 个神经元的 LSTM 网络,编解码层采用线性整流函数(Rectified Linear Unit,ReLU)为激活函数. 因为语言模型对最终的影响并不大[12],实验中并未使用语言模型. 采用的基线系统为局部注意力机制的语音识别模型,其编解码网络结构和所提出的方法均相同,唯一的区别是在注意力得分的计算方面. 在注意力得分计算部分,采用 5 个语音帧计算 $\hat{\alpha}_{i,j-1}$ 对 $\hat{\alpha}'_{i,j}$ 的约束因子. 在优化的前向注意力机制中,NN 模型采用一个含有 1024 个神经单元的隐含层,激活函数为 Sigmoid. 考虑到 fBank 在低信噪比的环境下依然能保证较好的鲁棒性[21],基音能够保证构建的语音字典的纯净性[22,23],本文采用 80 维的 fBank(Filter-Bank)和 3 维的基音(Pitch)作为输入每一帧的特征序列. 所有的模型通过 espnet[24,25]进行搭建,后端采用 pytorch 框架,在 Linux 系统的 Tesla K80 GPU 上运行.

5.3 实验结果及分析

在单头的注意力机制中,采用基于局部注意力模型("Att")的语音识别系统作为基线,在卷积滤波器尺寸分别为 25、50、100、200 时与前向注意力模型("For_Att")和优化前向注意力模型("ForTA_Att")进行对比. 对多头多尺度,采用 4 个多头且卷积滤波器尺寸都为 100 的模型("Mul_Head_Att")作为基线,多尺度模型("Mul_Scale_Att")每个头分别采用 25、50、100、200 的卷积窗构建不同的语音基元模型. 这里,卷积滤波器尺寸为 25 的模型主要代表着音素一级的模型,尺寸为 50、100 的模型主要代表着音节一级的模型,而尺寸为 200 的代表着词一级的模型. 本文还将前向注意力机制和多尺度模型进行结合,将每个头的传统注意力模型替换成前向注意力模型,从而构建出前向注意力机制多尺度模型("Mul_Scale_For_Att")和优化前向注意力机制多尺度模型("Mul_Scale_ForTA_Att"). 同时,为了证明前向注意力机制相比最近较为流行的语音识别算法更优,选择了近几年的方法作对比[9~12]. 具体结果见表 1 所示.

对于相关的研究工作,将文献[9]中没有使用语言模型和使用 LSTM 语言模型的注意力机制分别记为"Att_none"和"Att_LSTM". 它们相比于 Att 而言,词错误率相差不大,但相比于 For_Att 和 ForTA_Att,词错误率相对较高. 将文献[10]中对注意力机制采用了加窗约束的方法记为"Att_windowing",其词错误率依然高于本文提出的 For_Att 和 ForTA_Att 模型. 还将文献[11]中采用更复杂的编码模型的注意力机制记为"Att_2DLSTM",虽然在 SW 方面,其相比于 Att 取得了较好的结果,但在识别较为困难的 CH 数据集上词错误率比较高. 最后将文献[12]中不采用语言模型和采用 n-gram 和 RNN 作为语言模型的 CTC 方法分别记为"CTC_none"、"CTC_ngram"和"CTC_RNN". 由于这些模型存在帧独立性假设,因此 CTC 方法存在较高的词错误率.

表 1　模型测试词错误率和训练所花费的时间

模型	SW(%)	CH(%)	SW + CH(%)	time(h)
Att_200	13.3	24.8	19.4	141
Att_100	13.1	24.6	19.2	156
Att_50	12.8	24.1	18.7	177
Att_25	12.6	23.5	18.5	180
For_Att_200	12.8	24.6	19.2	145
For_Att_100	12.7	24.3	19	148
For_Att_50	12.5	23.8	18.6	151
For_Att_25	12.2	22.6	18.3	211
ForTA_Att_200	12.5	24.4	19	158
ForTA_Att_100	12.4	24.1	18.8	193
ForTA_Att_50	12.3	23.5	18.4	195
ForTA_Att_25	12.1	23	18.1	202
Mul_Head_Att	12.1	23.6	18.4	198
Mul_Scale_Att	11.9	23	18	204
Mul_Scale_For_Att	11.7	22.8	17.8	210
Mul_Scale_ForTA_Att	11.4	22.6	17.5	218
Att_none[9]	13.1	26.7	—	—
Att_LSTM[9]	11.8	25.7	—	—
Att_windowing[10]	16.2	29.1	22.7	—
Att_2DLSTM[11]	12.9	26.4	—	—
CTC_none[12]	24.7	37.1	—	—
CTC_ngram[12]	19.8	32.1	—	—
CTC_RNN[12]	14	25.3	—	—

从表 1 中能够看出:

(1)所提出的方法,在卷积滤波器尺寸都相等时,对当前时刻注意力得分进行平滑处理的 For_Att 优于传统 Att. 而在 For_Att 上加入约束的 ForTA_Att 的词错误率在 SW、CH、SW + CH 上相比于 For_Att 和 Att 都有明显降低.

(2)在 For_Att 和 ForTA_Att 中,随着滤波器尺寸的减少,模型对语音基元的建模逐渐细致,提取的语音信息也逐渐丰富,词错误率逐渐降低. 但由于较小的卷积滤波器对应着滑动窗的步长短,产生语音基元模型较多,所需要的计算量会相对更多,从而会导致模型训练所花费的时间较多.

(3)对多尺度的 Mul_Scale_Att 和单头的 For_Att 模型,相比于基线 Att,采用集成算法思想的 Mul_Scale_Att 识别效果提升比 For_Att 更加显著. 同时发现,Mul_Scale_Att 相比 Mul_Head_Att,前者对每个头进行音素、音节和词一级的语音基元建模,在训练时间上和 Mul_Head_Att 所花费的时间相差不大,但 WER 却得到了一定程度降低. 由(2)可知,对音素一级的语音基元进行建模能够在一定程度上提高模型的识别效果,以及对词一级的语音基元进行建模能够加快计算速度. 所以,Mul_Scale_Att 模型能在减少训练时间的基础上降低错误率.

(4)在 SW、CH、SW + CH 上相比于 Att_200,Mul_Scale_For_Att 的词错误率分别相对降低 12.03%、8.06%、8.25%,Mul_Scale_ForTA_Att 的词错误率分别相对降低 14.28%、8.87%、9.79%.

6　结论

针对传统注意力模型存在异常注意力得分,导致语音帧对齐效果差的问题. 本文首先提出前向注意力模型,其利用上一时刻正常的注意力得分对当前时刻可能存在异常的注意力得分进行平滑. 接着,通过对上一时刻注意力得分引入约束因子来自适应平滑异常得分. 最后,采用集成思想提出多尺度前向注意力模型,通过采用不同大小的卷积滤波器来对不同等级的语音基元进行建模,以挖掘更丰富的语音信息. 实验结果表明,多尺度前向注意力模型相比于基线系统 WER 相对降低 14.28%.

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [A]. Advances in Neural Information Processing Systems [C]. [S. l.]: NIPS, 2012. 1097 – 1105.

[2] CIRESAN D, GIUSTI A, et al. Deep neural networks segment neuronal membranes in electron microscopy images [A]. Advances in neural information processing systems [C]. [S. l.]: NIPS, 2012. 2843 – 2851.

[3] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(1): 30 – 42.

[4] DENG L, HINTON G, KINGSBURY B. New types of deep neural network learning for speech recognition and related applications: An overview [A]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing [C]. USA: IEEE, 2013. 8599 – 8603.

[5] COLLOBERT R, WESTON J. A unified architecture for natural language processing: Deep neural networks with multitask learning[A]. Proceedings of the 25th International Conference on Machine Learning [C]. USA: ACM, 2008. 160 – 167.

[6] HIRSCHBERG J, et al. Advances in natural language processing[J]. Science, 2015, 349(6245): 261 – 266.

[7] BAHDANAU D, CHOROWSKI J, et al. End-to-end atten-

tion-based large vocabulary speech recognition[A]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing[C]. USA:IEEE,2016. 4945 – 4949.
[8] GAO F C,XIN L I,YONG H Y. Using highway connections to enable deep small-footprint LSTM-RNNs for speech recognition[J]. Chinese Journal of Electronics,2019,28(1):107 – 112.
[9] ZEYER A,IRIE K,et al. Improved training of end-to-end attention models for speech recognition[A]. Interspeech[C]. Hyderabad:[s. n.],2018. 1845 – 1859.
[10] MERBOLDT A,ZEYER A,et al. An analysis of local monotonic attention variants[A]. Interspeech[C]. Graz:[s. n.],2019. 1398 – 1402.
[11] BAHAR P,ZEYER A,et al. On using 2D sequence-to-sequence models for speech recognition[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. USA:IEEE,2019. 5671 – 5675.
[12] ZWEIG G,et al. Advances in all-neural speech recognition[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. USA:IEEE,2017. 4805 – 4809.
[13] CHOROWSKI J,BAHDANAU D,CHO K,et al. End-to-end continuous speech recognition using attention-based recurrent nn:First results[J]. Eprint Arxiv,2014.
[14] CHAN W,et al. Listen,attend and spell:A neural network for large vocabulary conversational speech recognition[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. USA:IEEE,2016. 4960 – 4964.
[15] BAHDANAU D,CHOROWSKI J,et al. End-to-end attention-based large vocabulary speech recognition[A]. IEEE International Conference on Acoustics, Speech and Signal Processing[C]. USA:IEEE,2016. 4945 – 4949.
[16] MARTINS A,ASTUDILLO R. From softmax to sparsemax:A sparse model of attention and multi-label classification[A]. International Conference on Machine Learning[C]. [S. l.]:[s. n.],2016. 1614 – 1623.
[17] KIM Y. Convolutional neural networks for sentence classification[A]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)[C]. USA: Association for Computational Linguistics,2014. 1746 – 1751.
[18] SUTSKEVER I,VINYALS O,LE Q V. Sequence to sequence learning with neural networks[A]. Advances in Neural Information Processing Systems[C]. [S. l.]:NIPS,2014. 3104 – 3112.
[19] CHOROWSKI J K,et al. Attention-based models for speech recognition[A]. Advances in Neural Information Processing Systems[C]. [S. l.]:NIPS,2015. 577 – 585.
[20] GODFREY J J,HOLLIMAN E C,MCDANIEL J. SWITCHBOARD:Telephone speech corpus for research and development[A]. IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. USA:IEEE,1992. 517 – 520.
[21] WENBIN J,PEILIN L,FEI W. Speech magnitude spectrum reconstruction from MFCCs using deep neural network[J]. Chinese Journal of Electronics,2018,27(2):393 – 398.
[22] WEN M C,TIAN C H. The multi-weight neuron with geometry algorithm and its application[J]. Chinese Journal of Electronics,2008,17(2):261 – 264.
[23] JI X U,JIE L P,YONG H Y. Agglutinative language speech recognition using automatic allophone deriving[J]. Chinese Journal of Electronics,2016,25(2):134 – 139.
[24] KIM S,HORI T. Joint CTC-attention based end-to-end speech recognition using multi-task learning[A]. IEEE International Conference on Acoustics,Speech and Signal Processing[C]. USA:IEEE,2017. 4835 – 4839.
[25] WATANABE S,HORI T,KIM S,et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing,2017,11(8):1240 – 1253.

作者简介

唐海桃 男,1994 年 9 月出生于四川省广安市. 现为哈尔滨工业大学计算机科学与技术专业硕士研究生,主要研究方向为语音识别.
E-mail:tanghaitao_ape@163.com

薛嘉宾 男,1993 年7 月出生于内蒙古自治区包头市. 现为哈尔滨工业大学计算机科学与技术专业博士研究生,主要研究方向为语音识别.
E-mail:xuejiabin@hit.edu.cn

韩纪庆(通信作者) 男. 哈尔滨工业大学计算机科学与技术学院二级教授、学校长聘岗教授、博士生导师. 中国中文信息学会理事及语音处理专委会副主任、全国人机语音通讯学术会议常设机构委员会第二届、第三届主席团副主席、黑龙江省人工智能学会副理事长、黑龙江省计算机学会常务理事、《中文信息学报》编委、《数据采集与处理》杂志编委. 长期从事语音信号处理、音频信息处理等领域的教学与科研工作.
E-mail:jqhan@hit.edu.cn