文章编号:1003-0530(2018)06-0739-10

采用最少门单元结构的改进注意力声学模型

龙星延 屈 丹 张文林 徐思颖
(战略支援部队信息工程大学信息系统工程学院,河南郑州 450001)

摘 要:采用“编码-解码”结构的注意力声学模型存在参数规模庞大、收敛速度慢和在噪声环境中对齐关系不准确的问题。针对以上问题,先提出引入最少门结构单元减少模型参数,减少训练时间;再采用自适应宽度的窗函数和在计算注意力系数特征的卷积神经网络中加入池化层进一步提高音素与特征对齐的准确度,从而提升识别准确率。在英语和捷克语的实验结果表明,改进后的模型参数规模和音素错误率均下降,同时识别性能优于基于隐马可夫模型和基于连接时序分类算法的声学模型。

关键词:声学模型;注意力机制;最少门单元;自适应窗函数;池化层
中图分类号:TP912.3 **文献标识码**:A **DOI**:10.16798/j.issn.1003-0530.2018.06.013

An Improved Attention Based Acoustics Model with Minimal Gate Unit

LONG Xing-yan QU Dan ZHANG Wen-lin XU Si-ying
(Information System Engineering College, PLA Strategic Support Force Information Engineering University, Zhengzhou, Henan 450001, China)

Abstract: The acoustic model based “encoder-decoder” architecture with attention mechanism suffers from large scale, slow convergence and inaccurate distribution of attention due to the noise. In view of these problems, it is proposed to utilize Minimal Gate Unit to reduce the model parameters the training time. Then utilize the adaptive window function and add the pooling layer to the convolution neural network to improve the recognition accuracy as well as the accuracy of alignments between phonemes and acoustic features. The results of the experiments in English and Czech corpus show a certain decrease in quantity of parameters and the phone error rate, and the recognition performance outperforms the hidden Markov model based acoustic model and Connectionist Temporal Classification.

Key words: acoustic model; attention mechanism; minimal gate unit; adaptive window function; pooling layer

1 引言

声学模型(Acoustic Model, AM)是连续语音识别系统的核心模块,也是语音识别热门研究领域。由于隐马可夫模型(Hidden Markov Model, HMM)能描述语音信号时变性和非平稳性,同时拥有完成的理论体系和高效的模型参数估计与解码算法,它与高斯混合模型(Gaussian Mixture Model, GMM)组合成的 GMM-HMM 模型一直是主流的声学模型。伴随深度学习和人工智能技术等的兴起,深度神经网络(Deep Neural Network, DNN)与 HMM 组合的声学模型进一步提升识别率[1]。但基于 HMM 的声学模型存在以下缺陷:HMM 假设当前状态的先验概率只受上一状态影响,不能充分记录和利用音素序列的时序信息;HMM 建模依赖发音字典、决策树聚类等相关语言学知识。

为弥补 HMM 模型的缺陷,文献[2]提出在 GMM-HMM 框架上采用序列区分性准则重新训练模型,充分

收稿日期:2017-10-23;修回日期:2018-03-31
基金项目:国家自然科学基金(61673395, 61403415);河南省自然科学基金(162300410331)项目资助

学习特征序列的时序信息以提高识别准确率。在GMM-HMM框架下有效序列区分性准则包括最大互信息准则[3](Maximum Mutual Information, MMI)、增强型最大互信息准则[4](boosted MMI, bMMI)、最小音素错误[5](Minimum Phone Error, MPE)和最小贝叶斯风险[6](Minimum Bayes Risk, MBR)。文献[7]提出基于MMI准则的瓶颈深置信网络特征提取方法改进GMM-HMM系统性能。文献[8]在DNN-HMM模型中引入序列区分性准则,进一步提升声学模型的识别性能。Graves等人提出连接时序分类算法[9](Connectionist Temporal Classification, CTC),实现语音特征序列到音素序列的直接映射,建立基于CTC的端到端声学模型[10]。文献[11]在此基础上通过加权有限状态机将其与语言模型相结合并用于连续语音识别。与基于HMM声学模型相比,端到端模型不需要先验对齐信息和建立决策树等步骤,并且通过将字素作为建模对象可以摆脱对发音字典的依赖,但识别性能存在一定差距。

Cho等人提出一种基于循环神经网络的"编码-解码"端到端模型,并成功应用于机器翻译[12]。该模型通过编码网络将不同长度输入序列压缩成固定长度目标向量,解码网络再将目标向量作为特征识别逐一得到输出序列。Bahdanau等人在该模型中引入注意力机制,改进其在机器翻译任务中的性能[13]。注意力机制就是通过引入一个子网络对输入序列中所有元素进行关联度打分,再将归一化后的分数作为权重系数合成目标向量。注意力模型成功应用于图片自动标注[14]、音素识别[15]和连续语音识别[16]任务中。虽然该算法获得了性能的进一步提升,但仍然存在参数规模大、训练耗时极为严峻的问题,尽管通过硬件GPU可以部分解决,但从算法层面研究仍然是一个热点问题。此外,原始注意力声学模型存在在噪声环境下鲁棒性能差和音素与特征对齐不准确的问题[17]。

本文在基于注意力机制的"编码-解码"端到端模型基础上,提出了基于最少门单元结构的改进注意力声学模型。该模型首先将最少门结构单元替换原有循环神经网络单元,从而减少参数规模,提升训练速度;其次在计算注意力权重系数时,在文献[16]基础上,采用自适应宽度的窗函数和在计算注意力系数特征的卷积神经网络中添加池化层,进一步特征和音素对齐的准确度,进而提升声学模型的识别性能。

2 相关研究

2.1 基于门循环单元的循环神经网络

循环神经网络的内部呈环状结构,即当前时刻隐含层状态 $\boldsymbol{h}_t$ 可以表示以前一时刻隐含层状态 $\boldsymbol{h}_{t-1}$ 和当前时刻输入 $\boldsymbol{x}_t$ 为输入的函数,如式(1)所示。

$$\boldsymbol{h}_t = g(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \tag{1}$$

其中,g 为循环神经网络的传递函数,普通循环神经网络以式(2)作为传递函数。

$$g(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) = \boldsymbol{W}_{hx}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1} \tag{2}$$

其中,$\boldsymbol{W}_{hh}$ 为隐含层内部传递矩阵,$\boldsymbol{W}_{hx}$ 为输入层到隐含层传递矩阵。

最后将隐含层 $\boldsymbol{h}_t$ 作为输入,将sigmod函数作为激活函数计算得到输出层状态 $\boldsymbol{y}_t$,如式(3)所示。

$$\boldsymbol{y}_t = \mathrm{sigmod}(\boldsymbol{h}_t) \tag{3}$$

研究表明[18],由于普通RNN采用将隐含层状态与权重矩阵相乘的方式传递历史信息,导致训练过程中计算反向梯度时出现梯度消失和梯度爆炸的问题,无法有效传递长时记忆信息。为解决该问题,Hochreiter提出基于长短时记忆(Long Short-Term Memory, LSTM)单元的循环神经网络模型[18]。LSTM的传递函数 g 是一个复杂的非线性函数,内部设置记忆单元记录历史信息,通过门函数控制历史信息在特定时刻"累加"至隐含层状态,从而保证长时信息的有效传输。由于LSTM内部结构复杂,Cho提出门循环单元(Gate Recurrent Unit, GRU)。GRU结构只保留2个门函数且不包含额外记忆单元,在机器翻译测试中性能优于LSTM[12]。

给定特征序列 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$,GRU以式(4)作为传递函数得到隐含层序列 $(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$

$$\begin{aligned}
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_r),\\
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_z),\\
\boldsymbol{h}_t &= \boldsymbol{z}_t \odot \tanh(\boldsymbol{W}_h[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h) +\\
&\quad (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1}
\end{aligned} \tag{4}$$

其中,$\boldsymbol{r}_t$ 为重置门,$\boldsymbol{z}_t$ 为遗忘门,$\boldsymbol{W}_r$,$\boldsymbol{W}_z$,$\boldsymbol{W}_h$ 和 $\boldsymbol{b}_r$,$\boldsymbol{b}_z$,$\boldsymbol{b}_h$ 分别为权重矩阵和偏置向量,σ 和 tanh 为激活函

数。该传递函数可简记为式(5)。

$$\boldsymbol{h}_t = \mathrm{GRU}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \tag{5}$$

2.2 最少门单元

基于 GRU 单元的多层循环神经网络在应用时,需要保存和训练大量的参数,导致模型耗费大量存储空间,收敛速度较慢。针对该问题,采用文献[19]提出的最小门单元(Minimal Gated United, MGU)结构替代原始的 GRU 结构。MGU 结构具有更少的参数,并且在图像识别、语言模型、单词预测实验中的性能与 GRU 接近[19]。文献[19]中未在语音识别领域进行实验,本文将 MGU 结构应用于基于注意力的端到端声学模型,测试其在语音识别中的性能。

MGU 结构在 GRU 结构的基础上,让重置门 $\boldsymbol{r}_t$ 和遗忘门 $\boldsymbol{z}_t$ 共享一套参数,其传递函数如式(6)所示:

$$\begin{aligned} \boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_z) \\ \boldsymbol{h}_t &= \boldsymbol{z}_t \odot \tanh(\boldsymbol{W}_h[\boldsymbol{z}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h) + \\ &\quad (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} \end{aligned} \tag{6}$$

参数共享后,需要训练的权重矩阵从 $\boldsymbol{W}_h$、$\boldsymbol{W}_z$、$\boldsymbol{W}_r$ 减少至 $\boldsymbol{W}_h$、$\boldsymbol{W}_z$,使得参数规模减少 1/3。MGU 结构的传递函数可简记录为式(7):

$$\boldsymbol{h}_t = \mathrm{MGU}(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) \tag{7}$$

2.3 注意力机制

注意力机制是在序列到序列的模型中,通过模拟人类视觉机制,从输入特征序列中提取有效特征的技术。序列到序列模型中,需要先将变长特征序列$(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$映射成一个目标向量 $\boldsymbol{ct}$,该目标向量将序列中的重要信息进行压缩,从而实现变长序列到固定长度矢量的变换;再将 $\boldsymbol{ct}$ 作为输入,通过循环神经网络逐个计算出隐含层状态序列($\boldsymbol{s}_1$, $\boldsymbol{s}_2, \ldots, \boldsymbol{s}_O$),最终得到输出序列$(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_O)$。

文献[12]在机器翻译任务中将序列时刻 T 的特征 $\boldsymbol{h}_T$ 作为目标向量,即 $\boldsymbol{ct} = \boldsymbol{h}_T$。这种提取特征的方式没有利用特征序列其他时刻特征信息,因此表征能力受限。实际序列到序列建模问题中,例如机器翻译和语音识别,输出序列的元素总是与输入序列的特定元素存在对应关系,而采用注意力机制进行特征提取能够准确地描述和利用这种对应关系[13]。采用注意力机制计算输出序列位置 $o \in \{1, 2, \ldots, O\}$ 对应的目标向量 $\boldsymbol{ct}_o$ 过程如下:

首先,计算输出序列前一位置隐含层状态 $\boldsymbol{s}_{o-1}$ 与时刻 t 的特征的关联度,如式(8)所示:

$$e_{o,t} = a(\boldsymbol{s}_{o-1}, \boldsymbol{h}_t) \tag{8}$$

其中,$a(\cdot)$代表注意力子网络,它是只含一个隐含层的多层感知器,可表示式(9):

$$e_{o,t} = \boldsymbol{\omega}^{\mathrm{T}} \tanh(\boldsymbol{W}[\boldsymbol{s}_{o-1}, \boldsymbol{h}_t] + \boldsymbol{b}) \tag{9}$$

其中,$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 输入层到隐含层权重矩阵和偏置向量,$\boldsymbol{\omega}$ 隐含层到输出层权重矩阵。

然后,对所有时刻特征的关联度进行指数归一化。归一化后的数值称为注意力系数,如式(10)所示:

$$\alpha_{o,t} = \frac{\exp(e_{o,t})}{\sum_{t=1}^{T} \exp(e_{o,t})} \tag{10}$$

最后,将注意力系数作为权重,对所有时刻的特征加权求和,得到注意力机制下输出序列位置 o 的目标向量 $\boldsymbol{ct}_o$,如式(11) 所示:

$$\boldsymbol{ct}_o = \sum_{t=1}^{T} \alpha_{o,t} \boldsymbol{h}_t \tag{11}$$

采用注意力机制计算目标向量的过程,可以简记为式(12):

$$\boldsymbol{ct}_o = \mathrm{attention}([\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T], \boldsymbol{s}_{o-1}) \tag{12}$$

3 基于最少门单元的改进注意力声学模型

基于注意力机制的端到端模型最早应用于机器翻译[13],它能自动学习序列内部的时序信息,实现任意长度的输入序列到输出序列的直接建模。语音识别可看成是语音特征到音素的“翻译”,因此该模型也能应用于语音识别的声学模型。在基于注意力机制的端到端模型基础上,本文提出的改进算法模型如图 1 所示,模型由编码网络、解码网络和注意力子网络三个模块组成。编码网络采用基于 MGU 单元的深层循环神经网络,目的是学习和挖掘语音特征序列的前后依赖关系,提取语音的高层特征,增强特征的表达力和区分性;解码网络由基于 MGU 单元单层循环神经网络和 maxout 网络连接而成,目的是根据注意力机制得到的目标向量计算序列每个位置上所有音素出现的后验概率。注意力子网络是含一个隐含层的多层感知器,输入是上一时刻自身的输出,编码网络的输出,解码网络的前一个隐含层状态,输出是注意力系数。

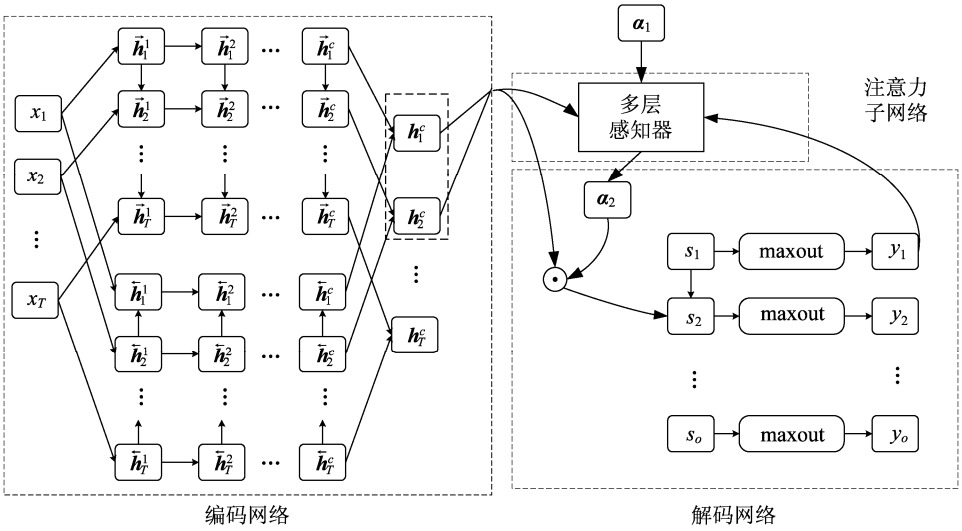

图 1 带注意力机制的“编码-解码”模型

Fig. 1 Structure of attention based encoder-decoder model

3.1 基于最少门单元的编码网络

基于最少门单元的编码网络中,将原有模型的 GRU 单元替换成 MGU 单元,目的是减少参数规模和训练时间。该编码网络由两组基于 MGU 单元的多层循环神经网络组成,分别为前向网络和后向网络。双向网络的结构能够同时传递过去和未来的信息,保证高层特征的信息量。网络输入为原始语音特征序列($\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T$),输出为高层特征序列($\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_T$)。

在时刻 t,第 c 层的前向网络和后向网络的隐含层状态分别为 $\vec{\boldsymbol{h}}_t^c$ 和 $\overleftarrow{\boldsymbol{h}}_t^c$,他们的隐含层单元信息传递方向相反,对应的传递函数分别为式(13)和式(14)。

$$\vec{\boldsymbol{h}}_t^1 = \mathrm{MGU}(\boldsymbol{x}_t,\vec{\boldsymbol{h}}_{t-1}^1) \tag{13}$$

$$\overleftarrow{\boldsymbol{h}}_t^1 = \mathrm{MGU}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}}_{t+1}^1) \tag{14}$$

前向网络和后向网络采用各自的传递函数并行进行层与层之间的特征传递,传递过程中在时域上进行降采样,以达到降低计算量的目标。以前向网络为例,将式(7)作为激活函数,由输入特征序列($\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T$)可得到第 1 层隐含层状态($\vec{\boldsymbol{h}}_1^1,\ldots,\vec{\boldsymbol{h}}_T^1$);同理,由 $c-1$ 层隐含层状态可计算出 c 层隐含层状态($\vec{\boldsymbol{h}}_t^c,\ldots,\vec{\boldsymbol{h}}_T^c$),计算过程隐含层状态如式(15)所示

$$\vec{\boldsymbol{h}}_t^c = \mathrm{MGU}(\vec{\boldsymbol{h}}_{2t}^{c-1},\vec{\boldsymbol{h}}_{2t-1}^c) \tag{15}$$

拼接前向网络和后向网络的第 c 层隐含状态,得到编码网络在时刻 t 的高层特征 $\boldsymbol{h}_t$,如式(16)所示:

$$\boldsymbol{h}_t = [\vec{\boldsymbol{h}}_t^c,\overleftarrow{\boldsymbol{h}}_t^c] \tag{16}$$

3.2 基于最少门单元的解码网络

解码网络由基于 MGU 的循环神经网络和 maxout 网络串联组成。它将编码网络计算得到的高层特征序列($\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_T$)作为输入,计算输出序列($\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_O$)。$\boldsymbol{y}_o$ 代表输出序列位置 o 上所有音素的后验概率,$\boldsymbol{y}_o$ 计算过程如下:

首先,解码网络将注意力子网络计算得到目标向量 $\boldsymbol{ct}_o$,作为基于 MGU 单元循环神经网络的输入,按照式(17)计算循环神经网络的隐含层状态 $\boldsymbol{s}_o$。

$$\boldsymbol{s}_o = \mathrm{MGU}(\boldsymbol{s}_{o-1},\boldsymbol{ct}_o) \tag{17}$$

然后,给定解码网络的隐含层状态 $\boldsymbol{s}_o \in R^d$ 作为输入条件下,通过 maxout 网络计算得到音素 i 的后验概率 $\boldsymbol{h}_i^{\mathrm{maxout}}$。maxout 网络的每个隐含层单元有多个候选单元,该网络从中选择数值最大的单元作为输出。计算过程如式(18)和式(19)所示:

$$h_i^{\mathrm{maxout}}(\boldsymbol{s}_o) = \max_{j\in[1,k]} z_{i,j} \tag{18}$$

$$z_{i,j} = \boldsymbol{s}_o^{\mathrm{T}}\boldsymbol{W}_{:,i,j} + \boldsymbol{b}_{i,j} \tag{19}$$

其中,d 为输入隐含层状态 $\boldsymbol{s}_o$ 的维度,对应隐含层单元数目,$\boldsymbol{W}_{:,i,j} \in R^{d\times m\times k}$ 和 $\boldsymbol{b}_{i,j} \in R^{m\times k}$ 为 maxout 网络参数矩阵和偏置向量,k 为 maxout 网络每个隐含层单元的候选单元数,m 为输出层单元数目,在声学模型中对应输出音素种类数目。

最后,如式(20)所示,由 maxout 网络的输出层得到输出向量 $\boldsymbol{y}_o$,$\boldsymbol{y}_o$ 第 i 个分量表示输出序列第 o 个位置上出息音素 i 后验概率

$$\boldsymbol{y}_o = [h_1^{\text{maxout}}, h_2^{\text{maxout}}, \ldots, h_m^{\text{maxout}}] \tag{20}$$

3.3　注意力机制改进

原始注意力模型中,注意力子网络对所有时刻的高层特征都计算关联度,而由于实际声学模型输出序列有很大概率出现相同音素,导致重复出现的音素在多个时刻的特征都拥有较大的关联度,从而造成注意力分散在错误的特征区域,影响识别性能。文献[16]通过增加窗函数限定注意力区域和增加卷积神经网络引入系数特征部分解决该问题,但仍然存在注意力对齐不准确地情况。在此基础上,我们采用自动调节窗口宽度的窗函数并且在卷积神经网络中加入池化层,进一步提升该模型在噪声环境中鲁棒性。采用自适应宽度的窗函数避免了注意力窗口内部注意力分布过于分散,并且减少音素对齐区域相重叠的现象,进而提升对齐关系的准确度。在卷积神经网络加入平均池化层能减轻噪声对注意力区域分布的干扰,从而增强模型鲁棒性。

3.3.1　自适应宽度的窗函数

在计算位置 o 音素后验概率时,窗口范围可表示为($m_o - w_{\text{L}}, \ldots, m_o + w_{\text{R}}$)。其中,$m_o$ 为窗口中心,w_{L} 为左窗长,w_{R} 为右窗长,对应窗函数取值(21)所示:

$$w_{o,i} = \begin{cases} 1, & m_o - w_{\text{L}} \leqslant i \leqslant m_o + w_{\text{R}} \\ 0, & \text{else} \end{cases} \tag{21}$$

限定范围后,每个时刻高层特征向量的关联度为:

$$\hat{e}_{o,i} = w_{o,t} e_{o,i} \tag{22}$$

把注意力系数 $\alpha_{o-1,t}$ 作为时刻 t 权重,则窗口中心 m_o 为加权平均数,如式(23)所示。

$$m_o = \sum_t t\alpha_{o-1,t} \tag{23}$$

注意力窗口宽度 w_{L},w_{R} 均固定设置为 50[16],原因是语音信号提取特征时帧移为 10 ms,100 帧宽的窗口能够利用 1 s 内的特征信息。一个音素的发音周期只有 0.2 ~ 0.5 s,所以限定范围后的注意力区域能完整覆盖 1 ~ 2 个音素的特征,能够保证注意力系数分布在正确的位置。

因为某些音素发音周期较短,固定长度的窗口音素数目可能较多,任然会出现相同的音素进而干扰注意力系数的分布,影响系统的识别性能。因此考虑设计能够根据前一个音素与特征对齐关系自动调整窗口宽度的窗函数。如图 2 所示,我们根据前两个音素的窗口中心 m_{o-1} 和 m_{o-2} 的偏移量估计出左窗长 w_{L},具体计算方式如式(24)所示。

$$w_{\text{L}} = \min\left(m_o - \frac{3m_{o-1} - m_{o-2}}{2}, w\right) \tag{24}$$

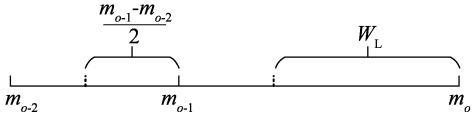

图 2　窗口宽度的计算

Fig. 2　Calculation of window width

3.3.2　计算系数卷积神经网络中增加池化层

使用卷积神经网络提取注意力系数特征时,将注意力系数向量 $\boldsymbol{\alpha}_{o-1}$ 先通过平均池化层,再通过一维卷积层,目的是提升系数特征鲁棒性和区分性。输入的注意力系数向量为前一个音素对应层窗口内所有注意力系数,如式(25)所示。

$$\boldsymbol{\alpha}_{o-1} = [\alpha_{o-1,m_{o-1}-w}, \ldots, \alpha_{o-1,m_{o-1}+w}] \tag{25}$$

其中,由于采用的是自适应窗长,所以令 $\alpha_{o-1,m_{o-1}-w} = \alpha_{o-1,m_{o-1}-w+1} = \cdots = \alpha_{o-1,m_{o-1}-w_{\text{L}}} = 0$。

池化层采用平均池化,池化滤波器的规模为 1×3,池化前后向量维度保持不变。卷积层的卷积核大小为 1×(2w+1),滤波器数目为 j,卷积方式采用 same padding。以上卷积过程可以由式(26)、(27)描述:

$$\bar{\alpha}_{o-1,t} = \frac{\alpha_{o-1,t-1} + \alpha_{o-1,t} + \alpha_{o-1,t+1}}{3} \tag{26}$$

$$\boldsymbol{l}_o = \boldsymbol{L} \otimes \bar{\boldsymbol{\alpha}}_{o-1} \tag{27}$$

其中,$\bar{\boldsymbol{\alpha}}_{o-1}$ 为池化层后注意力系数 $\bar{\alpha}_{o-1,t}$($t \in [m_{o-1} - w, m_{o-1} + w]$)组成的 2$w$+1 维向量,$\boldsymbol{L}$ 为 j×(2w+1)矩阵,卷积后矩阵 $\boldsymbol{l}_o$ 规模也为 j×(2w+1)。

最后将卷积神经网络的输出用于计算注意力系数得分,公式如式(27)所示

$$e_{o,t} = \boldsymbol{\omega}^{\text{T}} \tanh(\boldsymbol{W}[\boldsymbol{s}_{o-1}, \boldsymbol{h}_t, \boldsymbol{l}_{o,t}] + \boldsymbol{b}) \tag{28}$$

其中，$\boldsymbol{l}_{o,t}$为矩阵 $\boldsymbol{l}_o$ 中时刻 t 对应的列向量。

3.4 模型训练和解码

虽然模型中注意力子网络和解码网络的连接呈现环状结构，但依然通过最优化目标函数的方式训练模型参数。解码时由于输出序列长度未知，需要采用带序列终止符的 BeamSearch 算法解码。

对于含 N 段语音的数据集，模型训练采用梯度下降法求目标函数的最小值，目标函数如式(29)所示

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N}\sum_{n=1}^{N} -\log P(p_1^n,\ldots,p_{O_n}^n \mid \boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_{T_n}^n,\theta) \tag{29}$$

其中，n 为语音的编号，$p_1^n,\ldots,p_o^n$为该语音的正确音素标注序列，$\boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_T^n$ 为该语音的特征序列，θ 为模型的全部参数，即编码网络、解码网络和注意力子网络中所有的权重矩阵和偏置向量。

模型由于是对序列进行建模，所以单段语音后验概率的计算方式如式(30)所示：

$$\begin{aligned} P(p_1^n,\ldots,p_o^n \mid \boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_T^n,\theta) &= P(p_1^n \mid \boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_{T_n}^n,\theta) \times \\ P(p_2^n \mid \boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_{T_n}^n,\theta) &\times \cdots \times P(p_O^n \mid \boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_{T_n}^n,\theta) \\ &= \boldsymbol{y}_1^{p_1} \times \boldsymbol{y}_2^{p_2} \times \cdots \times \boldsymbol{y}_{O_n}^{p_O} \end{aligned} \tag{30}$$

其中，$y_o^{p_o}, o \in \{1,2,\ldots,O\}$代表模型的解码网络生成输出序列中第 o 个输出向量对应音素 p_o 的后验概率。

解码带序列终止符的 BeamSearch 算法搜索在解码网络的输出中寻找负概率值最低的序列作为输出。该算法的思想是维护一个容量为 beam_size 序列集合，每步搜索时将集合中的序列拓展一位，然后筛选结果最好的 beam_size 个序列保留在集合中。具体流程见算法 1。

其中，tmp 为拓展后的序列集合，beam 为算法维护的候选搜索序列集合，done 为含有终止符<eos>的序列集合，phone 为解码网络在位置上音素对应的后验概率，phone_set 为所有音素和终止符<eos>的集合，best 记录最低的负概率值，m_length 为序列长度上限。

算法 1 带终止符的 BeamSearch 解码算法

输入 每个位置的音素后验概率 $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_O$

输出 负概率值最低的音素序列 $p_1,\ldots,p_o$

初始化：beam = {∅}，tmp = ∅，done = ∅，best = 1e9，m_length = 1e4

1. while beam = {∅} and $i \leqslant$ m_length
2. $i = i+1$, tmp = ∅
3. for sequence in beam
4. for phone in phone_set
5. newsequence. list = sequence. list + phone
6. newsequence. cost = sequence. cost + y_i^{phone}
7. tmp = tmp + newsequence
8. beam = ∅
9. 筛选出集合 tmp 中以<eos>结束的序列加入集合 done。
10. 对集合中 done 序列按照 cost 升序排序。
11. 如果 done[1]. cost < min_cost 则 best = done[1]. cost。
12. 如果连续 50 步 best 未更新，则跳出 while 循环。
13. 对 tmp 中的序列按照 cost 升序排序。
14. 将 tmp 中 cost 最小的 beam_size 个序列加入集合 beam。
15. end while
16. 将 done[1]. list 音素序列作为结果输出。

4 实验

为了验证本文改进方法的有效性，并与 HMM 声学模型和链接时序分类方法进行对比，我们采用了语音识别测试中常用的英语语料库和捷克语语料库作为数据集。特征提取采用 kaldi[20] 开源工具包，端到端模型基线系统采用 Theano[21] 开源深度学习库搭建。

4.1 实验数据

TIMIT 语料库是语音识别领域最常用的标准数据库之一，它包含 6300 段英语朗读语音。在实验中选取 3296 条语句作为训练集，192 条语句作为测试集，400 条语句作为开发集。

Vystadial_cz 是开源捷克语语料库，它包含 15 小时电话信道下的含噪声对话语音，识别率普遍较低。训练集有 22566 条语句，测试集和开发集各有 2000 条语句。

4.2 实验设置

特征提取:语音信号采样频率是 16 kHz,采样位 16 bit,使用 Hamming 窗处理,帧长 25 ms,帧移 10 ms,预加重系数 0.97。语音输入特征向量采用 40 维 fbank 特征和能量,再拼接对应一阶和二阶差分,共计 123 维参数。对于提取好的特征,首先在训练集范围内进行归一化,使每个分量服从标准正态分布,再利用训练集的归一化参数对测试集和开发集特征归一化处理。

模型初始化:循环神经网络权重矩阵初始设定为标准正交矩阵,偏置向量初始设为 0,内部状态值采用均值为 0 方差为 0.1 的独立高斯分布初始化。

模型参数:编码网络的隐含层状态维度设为 200。注意力子网络的卷积神经网络通道数设为 10。英语声学模型解码网络输出向量设为 63 维,分别对应 61 个音素、空白符和序列终止符的后验概率;捷克语声学模型解码网络输出向量设为 44 维,分别对应 41 个捷克语字母,空白符、噪声符号和序列终止符出现的概率。maxout 网络的候选隐含层数目设为 64。

模型训练:以式(29)作为目标函数,使用随机梯度下降法(Stochastic Gradient Descent, SGD)对模型参数迭代更新。训练过程分为两个阶段:第一阶段样本批量大小(batch size)为 8,目的是高训练效率,使模型参数尽快收敛;第二阶段样本批量大小为 1,每次训练时给模型添加噪声,目的是增强模型识别的鲁棒性和抗噪能力。

4.3 评价指标

TIMIT 数据集的识别结果为音素序列,考虑采用动态规划算法将模型解码得到的序列与标注序列以音素作为基本单元对比并统计出插入错误(I)、删除错误(D)和替代错误(R)。设测试集中含有 N 个句子,则音素错误率(Phone Error Rate, PER)为:

$$\mathrm{PER} = \frac{\mathrm{I} + \mathrm{D} + \mathrm{R}}{N} \times 100\% \qquad (31)$$

Vystadial_cz 数据集的识别结果为识别结果捷克字母序列,将字母序列整合成单词,并以单词作为基本单元统计出词错误率(Word Error Rate, WER)作为声学模型评价指标。

为评价和对比注意力模型训练速度,将训练过程的第二阶段中批量大小为 1 条件下,用训练集的所有样本更新模型参数的平均周期(epoch)作为评价指标。

4.4 实验结果和分析

(1)基于 MGU 单元和 GRU 单元系统性能对比

为了对比采用不同单元的系统性能影响,在标准 TIMIT 语料库中进行音素序列识别的实验。表 1 给出在 TIMIT 测试集中,编码网络中循环神经网络为 1 至 3 层时,分别采用 GRU 和 MGU 作为循环神经网络基本单元时的性能。由表 1 可以看出,对于同样的模型结构,增加隐含层层数数目,系统的识别性能得到提升,但系统的参数规模和训练周期也迅速上升。在相同层数下,MGU 的参数规模和平均迭代周期均低于 GRU。2.2 中提到隐含层维度相同的 MGU 的参数规模为 GRU 的 2/3,因此当编码网络层数增加 1 层时,GRU 结构参数的增长规模是 MGU 结构的 1.5 倍。当编码网络层数为 3 层时,MGU 的参数规模下降 39.0%,平均迭代周期下降 14.7%,而测试集的音素错误率仅高 0.1%。以上实验结果证明在基于注意力的端到端声学模型中,使用 MGU 替换 GRU 能够在识别性能损失较小的前提下,有效减少参数规模和提高收敛速度。

(2) 改进注意力机制前后系统性能对比

为验证使用 3.2 中改进注意力机制方法的有效性,分别在 TIMIT 和 Vystadial_cz 语料库搭建声学模型进行实验。采用 3 层 GRU 结构的编码网络作为基线系统,先采用固定长度的窗口和无池化层的卷积神经网络对注意力机制进行优化,窗口宽度为 100,卷积神经网络的滤波器数目设置为 10;再分别采用自适应宽度的窗函数和加入池化层的卷积神经网络方法。表 2 的结果表明,对于 TIMIT 语料库上,改进后的模型在开发集上准确率提升明显,在测试集上的准确率提升较小。对于噪声较大的 Vystadial_cz 语料库,采用自适应宽度的窗函数和增加池化层后对于测试集的音素错误率与改进前模型相比下降 1.06% 和 0.68%。这证明改进后的注意力模型能够更准确地计算音素和编码网络特征的关联度,拥有更好识别性能和对噪声有更强鲁棒性。

表1 TIMIT语料库不同模型的性能

Tab. 1 Performance of different models in TIMIT datasets

层数	模型结构	开发集错误率/%	测试集错误率/%	参数规模/M	平均迭代周期/min
1	GRU	20.65	21.83	1.986	36
	MGU	20.24	21.66	1.530	30
2	GRU	17.49	21.14	2.570	55
	MGU	18.67	20.39	1.920	47
3	GRU	**17.06**	**19.57**	3.155	75
	MGU	17.23	19.67	2.310	64

表2 TIMIT和Vystadial_cz语料库下不同系统的音素(词)错误率

Tab. 2 PER(WER) of different systems in TIMIT and Vystadial_cz datasets

数据集	系统	开发集错误率/%	测试集错误率/%
TIMIT	基线系统	17.0	19.57
	固定窗函数+无池化层	16.87	18.06
	自适应窗函数+无池化层	**16.42**	17.91
	固定窗函数+池化层	16.52	18.03
	自适应窗函数+池化层	16.51	**17.90**
Vystadial_cz	基线系统	61.12	60.33
	固定窗函数+无池化层	55.41	54.53
	自适应窗函数+无池化层	52.97	53.47
	固定窗函数+池化层	54.46	53.85
	自适应窗函数+池化层	**52.27**	**53.20**

为了更加直观地改进前后系统的变化,提取Vystadial_cz语料库中一段语音,打印出基线系统与采用“固定窗函数+无池化层”与“自适应窗函数+池化层”两个模型识别出的音素与特征的对齐情况,如图3和图4所示。图中竖轴代表音素序列,横轴表示高层特征帧数,色块颜色深浅表示注意力系数大小。由于编码网络对特征进行降采样处理,横轴的显示帧数是实际语音帧数的四分之一。通过对比观察可以得到,图3中捷克字母S、L的注意力系数在距离窗口中心较远区域仍有分布,而图4中注意力系数分布更为精确。

(3)端到端声学模型与其他模型对比

本文对比了改进前后基于注意力机制的端到端声学模型与其他声学模型在无语言模型条件下连续语音识别任务中的性能,以验证该模型和改进方法的有效性。这里涉及到的声学模型包括:基于三音子的GMM-HMM模型,采用MMI优化GMM-HMM模型,采用bMMI优化GMM-HMM模型,采用MPE优化GMM-HMM模型,采用sMBR优化的DNN-HMM模型和基于RNN-CTC模型。基于注意力的端到端声学模型中分别采用基线系统,改进注意力机制后的模型以及替换MGU单元后的改进模型。

表3给出了本文改进模型与其他模型在Vystadial_cz数据集上的实验结果对比。由表中可以得出,在传统方法中,采用深度神经网络和sMBR准则优化的声学模型性能最佳,在测试集的性能由于注意力模型额基线系统。基于RNN-CTC的端到端声学模型虽然不依赖发音字典等先验知识,但在该数据集下识别性能不如传统方法。改进注意力的端到端声学模型在开发集和测试集性能最佳,原因是它能更加充分地学习和利用语音中时序信息,并且能让音素和特征更加准确地对齐。将改进系统的GRU单元替换成MGU单元后,虽然为了减少模型参数规模和提升收敛速度损失了少部分识别性能,但词错误率依然低于其他声学模型。

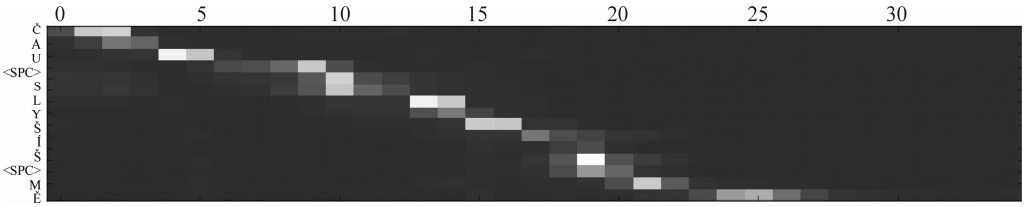

图 3　改进前系统音素与特征对齐情况

Fig. 3　Alignment of unimproved system between phones and features

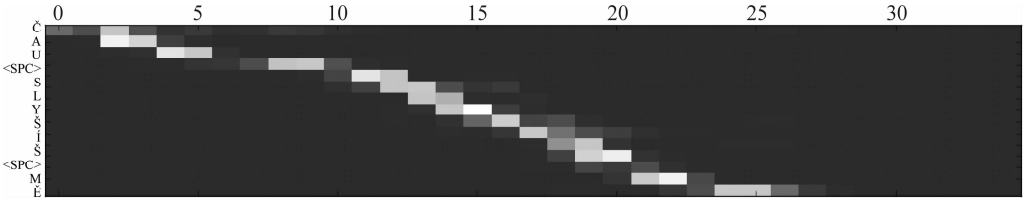

图 4　改进后系统的音素与特征对齐情况

Fig. 4　Alignment of improved system between phones and features

表 3　Vystadial_cz 语料库下各个系统的词错误率

Tab. 3　WER of different contious speech recogniton system

系统	开发集错误率/%	测试集错误率/%
GMM-HMM	79.11	65.87
GMM-HMM MMI[3]	76.08	67.38
GMM-HMM bMMI[4]	75.54	66.94
GMM-HMM MPE[5]	73.97	63.08
DNN-HMM sMBR[8]	69.56	57.74
RNN-CTC[10]	77.72	69.05
注意力模型	55.41	54.53
注意力改进模型	**52.27**	**53.20**
注意力改进模型(MGU)	53.64	53.32

5　结论

本文研究了基于注意力机制的端到端声学模型。在基线系统的基础上,先采用 MGU 替代 GRU 作为循环神经网络基本单元,在损失识别率较低情况下,降低了模型参数规模和训练时间。再根据语音信号特点通过使用自适应宽度的窗函数和在计算注意力系数特征的卷积神经网络中加入池化层,进一步提高了模型的识别准确率。在捷克语语料库下的实验表明,改进后模型的识别率优于基于 HMM 声学模型和基于 CTC 的端到端模型。下一步的研究方向是寻找更高效的提取语音特征方法,调整系统内部结构和训练准则以降低训练复杂度,提升识别性能。

参考文献

[1] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6):82-97.

[2] Bahl L, Brown P, De Souza P, et al. Maximum mutual information estimation of hidden markov model parameters for speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 1986:49-52.

[3] Valtchev V, Odell J, Woodland P C, et al. Lattice-based discriminative training for large vocabulary speech recognition[C]// Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings. 1996 IEEE International Conference. IEEE Computer Society, 1996:605-608.

[4] Povey D, Kanevsky D, Kingsbury B, et al. Boosted MMI for model and feature-space discriminative training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008:4057-4060.

[5] Povey D, Woodland P C. Minimum Phone Error and I-smoothing for improved discriminative training[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2002:I-105-I-108.

[6] Povey D, Kingsbury B. Evaluation of Proposed Modifications to MPE for Large Scale Discriminative Training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2007:IV-321-IV-324.

[7] 陈雷, 杨俊安, 王一, 等. LVCSR 系统中一种基于区分性和自适应瓶颈深度置信网络的特征提取方法[J]. 信号处理, 2015, 31(3):290-298.

Chen Lei, Yang Junan, Wang Yi, et al. A Feature Extraction Method Based on Discriminative and Adaptive Bottleneck Deep Belief Network in Large Vocabulary[J]. Journal of Signal Processing, 2015, 31(3):290-298. (in Chinese)

[8] Voigtlaender P, Doetsch P, Wiesler S, et al. Sequence-discriminative training of recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015:2100-2104.

[9] Graves A. Connectionist Temporal Classification[M]// Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, 2012:61-93.

[10] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]// International Conference on Machine Learning, 2014:1764-1772.

[11] Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]// Automatic Speech Recognition and Understanding. IEEE, 2016:167-174.

[12] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[13] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[14] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [J]. Computer Science, 2015.

[15] Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-Based Models for Speech Recognition[J]. Computer Science, 2015.

[16] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016:4945-4949.

[17] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017:4835-4839.

[18] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8):1735-1762.

[19] Zhou G B, Wu J, Zhang C L, et al. Minimal Gated Unit for Recurrent Neural Networks[J]. International Journal of Automation and Computing, 2016, 13(3):226-234.

[20] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[EB/OL]. http://www.danielpovey.com/files/2011_asru_kaldi.pdf. 2011.

[21] Bergstra J, Breuleux O, Bastien F, et al. Theano: a CPU and GPU math expression compiler[EB/OL]. http://conference.scipy.org/scipy2010/slides/james_bergstra_theano.pdf. 2010.

作者简介

龙星延 男, 1992 年生,广西柳州人,战略支援部队信息工程大学硕士研究生,主要研究方向为语音识别、机器学习与人工智能。

E-mail:lxy120999@qq.com

屈 丹(通信作者) 女,吉林九台人。战略支援部队信息工程大学副教授,主要研究方向为语音识别、智能信息处理。

E-mail:qudanqudan@163.com

张文林 男,1982 年生,湖北人。战略支援部队信息工程大学副教授,主要研究方向为语音识别、机器学习与人工智能。

E-mail:zwlin_2004@163.com

徐思颖 女,1992 年生,河南南阳人。战略支援部队信息工程大学硕士研究生,主要研究方向为语音增强、智能信息处理。

E-mail:xusiying_2015@163.com