融合语言模型的端到端中文语音识别算法

吕坤儒[1],吴春国[1,2,3],梁艳春[1,2,3],袁宇平[1],任智敏[1],周 柚[1,2],时小虎[1,2,3]
(1. 吉林大学计算机科学与技术学院,吉林长春 130012; 2. 吉林大学符号计算与知识工程教育部重点实验室,吉林长春 130012;
3. 珠海科技学院计算机学院,广东珠海 519041)

摘 要: 为了解决语音识别模型在识别中文语音时鲁棒性差,缺少语言建模能力而无法有效区分同音字或近音字的不足,本文提出了融合语言模型的端到端中文语音识别算法.算法建立了一个基于深度全序列卷积神经网络和联结时序分类的从语音到拼音的语音识别声学模型,并借鉴Transformer的编码模型,构建了从拼音到汉字的语言模型,之后通过设计语音帧分解模型将声学模型的输出和语言模型的输入相连接,克服了语言模型误差梯度无法传递给声学模型的难点,实现了声学模型和语言模型的联合训练.为验证本文方法,在实际数据集上进行了测试.实验结果表明,语言模型的引入将算法的字错误率降低了21%,端到端的联合训练算法起到了关键作用,其对算法的影响达到了43%.和已有5种主流算法进行比较的结果表明本文方法的误差明显低于其他5种对比模型,与结果最好的DeepSpeech2模型相比字错误率降低了28%.

关键词: 语音识别; 联结时序分类; 语言模型; 声学模型; 语音帧分解
中图分类号: TP18;TP39 **文献标识码:** A **文章编号:** 0372-2112(2021)11-2177-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20201187

An End-to-End Chinese Speech Recognition Algorithm Integrating Language Model

LÜ Kun-ru[1], WU Chun-guo[1,2,3], LIANG Yan-chun[1,2,3], YUAN Yu-ping[1], REN Zhi-min[1],
ZHOU You[1,2], SHI Xiao-hu[1,2,3]
(1. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Ministry of Education Key Laboratory of Symbol Computation and Knowledge Engineering, Jilin University, Changchun, Jilin* 130012, *China*;
3. *School of Computer Science, Zhuhai College of Science and Technology, Zhuhai, Guangdong* 519041, *China*)

Abstract: To address the problems of poor robustness, lack of language modeling ability and inability to distinguish between homophones or near-tone characters effectively in the recognition of Chinese speech, an end-to-end Chinese speech recognition algorithm integrating language model is proposed. Firstly, an acoustic model from speech to Pinyin is established based on Deep Fully Convolutional Neural Network (DFCNN) and Connectionist Temporal Classification (CTC). Then the language model from Pinyin to Chinese character is constructed by using the encoder of Transformer. Finally, the speech frame decomposition model is designed to link the output of the acoustic model with the input of the language model, which overcomes the difficulty that the gradient of loss function cannot be passed from the language model to the acoustic model, and realizes the end-to-end training of the acoustic model and the language model. Real data sets are applied to verify the proposed method. Experimental results show that the introduction of language model reduces the word error rate (WER) of the algorithm by 21%, and the end-to-end integrating training algorithm plays a key role, which improves the performance by 43%. Compared with five up-to-date algorithms, our method achieves a 28% WER, lower than that of the best model among comparison methods—DeepSpeech2.

Key words: speech recognition; CTC; language model; acoustic model; speech frame decomposition

收稿日期:2020-10-23;修回日期:2021-07-20;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61972174);吉林省预算内基本建设资金(No.2021C044-1);广东省国际科技合作项目(No.2020A0505100018);吉林省自然科学基金(No.20200201163JC)

1 引言

作为实现和改善智能人机交互的重要技术,语音识别技术在过去的几十年里一直都是研究热点[1]. 在深度学习技术兴起之前,语音识别技术主要是基于高斯模型和隐马尔可夫模型的混合模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)[2]. 随着深度学习技术的兴起, GMM逐渐被建模能力更强大的深度神经网络(Deep Neural Network, DNN)所替代[3,4],语音识别框架变为DNN-HMM. 之后递归神经网络(Recurrent Neural Network, RNN)、长短时记忆网络(Long Short-Term Memory, LSTM)和卷积神经网络(Convolutional Neural Network, CNN)也逐渐被应用到语音识别领域[5~7]. 但这个时期的语音识别框架本质上还是以HMM为核心模型,而HMM模型的多种不合理假设存在诸多弊端. 为了简化语音识别系统联合多模块共同优化,联结时序分类算法(Connectionist Temporal Classification, CTC)[8]被引入到语音识别领域,实现了语音序列和文字序列的自动对齐,语音识别进入端到端时代. 例如Pezeshki等[9]将更深的CNN与CTC结合;使用SpecAugment数据增强算法,Yang等[10]提出了一个基于Transformer的端到端模型,提高了在藏语上的识别率;Chang等结合分频CNN特征提取器,获得了在耳语数据集上效果理想的端到端模型[11];由于数据增强算法受到失真数据的影响, Fan团队提出递归融合方法有效去除噪声信号,减轻语音失真问题[12];Graves提出使用双向LSTM网络来对当前帧进行处理[13];Dinkel等[14]以原始语音波形作为输入并将LSTM和CNN进行结合提出CLDNN混合架构;同时百度也提出将双向GRU和CNN结合构造网络层数更深的Deep Speech2模型[15]并在中英文混合语料中取得了不错的识别率;其中最具代表性的是科大讯飞提出的深度全卷积神经网络(Deep Fully Convolutional Neural Network, DFCNN)[16],网络中全部使用卷积操作直接对语音信号进行建模,通过积累非常深的卷积池化操作使得模型能学习到更多的历史信息,实验表明DFCNN比BLSTM语音识别系统这个学术界和工业界最好的系统识别率提升了15%以上.

目前大多数方法都是针对声学模型展开工作,未考虑语言模型在训练过程中对声学模型的影响. 融合声学模型和语言模型的最常用的方法是浅层融合(Shallow Fusion),主要做法是分别训练声学模型和语言模型,然后组合它们的输出以引导束集搜索,外部语言模型只参与声学模型的解码过程[17]. Bengio[18]在机器翻译任务上提出深层融合(Deep Fusion)为语言模型的融合带来了新的思路. 此外,Sriram等人[19]在深层融合的基础上提出冷融合(Cold Fusion). Toshniwal等人[20]针对这三种不同思路的融合方法分别在VS14k和D15K上进行评估,冷融合要比深层融合效果好,但是后两种方法无论是在模型效果还是在简洁度上都不如浅层融合,且深层融合和冷融合方法并未与当前主流端到端模型CTC结合使用,因此语音识别领域中主流仍然是浅层融合. 此外,以上的融合方法使用的语言模型主要是N-gram统计模型. 随着深度学习的发展,找到优化的策略或网络结构对语言模型进行改进也颇具研究价值.

本文针对CTC端到端模型缺少外部语言模型以及无法联合优化的难点问题进行研究,在DFCNN模型的基础上,针对汉语语言的发音特性提出新的建模单元集合,建立基于Transformer编码器结构的语言模型,并将其与声学模型进行深度融合,通过设计语音分帧模块克服了语言模型和声学模型无法联合训练的难点,提出了融合语言模型的端到端语音识别算法(End-to-end Chinese Speech Recognition algorithm Integrating Language Model, ECSRILM). 在一定程度上纠正了基于CTC语音识别系统产生的同音字或近音字的替换错误,整体上提高了识别正确率. 最后通过实际数据集对上述端到端模型有效性进行验证,模型最终错误率降至11. 88%.

2 联结时序分类算法

本文在声学模型设计上摒弃了传统的隐马尔可夫模型,引入联结时序分类CTC算法并将其应用于声学模型的训练,自动完成序列对齐任务.

联结时序分类CTC是由Graves等提出的用来解决时序类数据分类的方法. CTC与传统的基于HMM的声学模型不同,不需要对数据进行帧级别的强制对齐,而是通过在输出序列中加入空白标签Blank来实现语音帧序列与文本序列的自动对齐,该过程极大地简化了端到端模型的训练流程.

CTC训练是在网络输出层应用CTC目标函数,自动完成输入序列与输出标签之间的对齐. 对于序列标记问题,假设给定语音输入序列$X=(x_1,x_2,\cdots,x_T)$和对应输出序列$Y=(y_1,y_2,\cdots,y_U)$, 其中T为语音的时间窗长度,x_i为在第i个时间窗所对应的语音特征向量;U为输出音节的个数,$y_i\in L$为输出的第i个音节,L是输出序列集合. CTC训练的目标就是在给定输入序列X下,通过调整模型参数最大化输出标签序列的对数概率即$\max(\ln P(Y|X))$. 为了解决可变序列的对齐问题,首先扩充输出序列集合$L'=L\cup\{\text{blank}\}$,之后引入一个与输入序列在帧上一一对应的CTC路径$\pi=(\pi_1,\pi_2,\cdots,\pi_T)$,在CTC路径中允许空白标签和非空白标签连续重复出现,即$\pi_i\in L'$,最后的目标序列Y由路径π中相邻的重复字符合并,再删除空字符后得到. 假定每一时刻的输出之间条件独立,整个CTC路径的概率可以由每一帧对应标签的概率组合而成:

$$P(\pi|X)=\prod_{t=1}^{T}P(\pi_t|X_t) \tag{1}$$

对于目标序列Y可以由多个CTC输出序列π与之对应,因此可以用所有CTC路径的概率来表示输出标签Y的概率:

$$P(Y|X)=\sum\nolimits_{p\in\beta(Y)}P(p|X) \tag{2}$$

其中β是从π到Y的映射,该映射先合并相邻重复出现的类,再去除空类. 由于CTC路径的所有可能情况会随输入序列规模呈指数式增长,导致计算复杂度太大,所以上式可通过动态规划算法中的前向后向算法在篱笆网络中高效地计算路径似然度. 因此设定CTC损失函数如式所示:

$$CTC(X)=-\ln P(p|X)=-\sum\nolimits_{p\in\beta(Y)}\prod_{t=1}^{T}y_t^{p'} \tag{3}$$

经过训练之后的网络即可应用于语音样本的预测,CTC最终输出一个$T\times N$的概率矩阵,其中T为输入序列长度,N为分类器的类别数,该矩阵可通过特定的搜索算法如贪心搜索(Greedy-search)和集束搜索(Beam-search)来找出概率最大的声学单元序列.

3　本文方法

传统的语音识别系统在识别阶段通常会联合声学模型和语言模型进行解码,以充分利用外部语言模型的语言学知识. 一方面由于CTC模型输出独立无关的假设,认为每个时刻的预测样本之间是无关的忽略了语音信息之间的相关性,所以如果能够在CTC中加入语言模型就可以改善这一不合理假设带来的影响. 另一方面,对于中文语言来说由于同音异义字的存在,纯靠声学模型往往很难对其有效区分,因此需要联合语言模型利用文本的语义信息加以补充. 针对CTC端到端语音识别模型中缺少语言建模能力以及不能有效整合语言模型进行联合优化等不足,我们提出了一种新的融合语言模型的端到端语音识别模型(End-to-end Chinese Speech Recognition algorithm Integrating Language Model, ECSRILM),即包含Transformer语言模型的深度全卷积神经网络(DFCNN). 通过设计语音分帧模块建立了声学模型输出与语言模型输入的有效连接,从而实现了语言模型和声学模型的协同训练. 本节首先介绍一下模型的整体框架,之后对声学模型、语音帧分解模型和语言模型三个主要算法模块进行详细阐述.

3.1　模型框架

本文所提出的ECSRILM的基本框架是将声音信号转为语谱图之后采用基于DFCNN框架的声学模型转化为音节序列,进而作为基于Transformer编码器结构的语言模型的输入,得到最终的汉字序列. 但是因为音节序列与语谱图的语音帧并不等长,所以无法直接输入到语言模型中,因此设计了语音帧分解模型进行匹配. 在模型的训练过程中采用了迁移学习和微调的思想,即首先分别对声学模型和语言模型进行预训练,之后将两者进行深层融合,通过计算语音分帧权重矩阵来实现两者的连接,从而对整体模型联合进行微调训练. 模型结构框架如图1所示. 首先要将语音信息按时间窗进行分帧处理,并转化为类似于图像的语谱图,将其作为声学模型DFCNN的输入. 在预训练过程中,声学模型的输出为汉语拼音,而在联合训练的整体框架中,需要截掉最后的CTC解码层,将Dense(256)层作为输出,并与语音分帧矩阵相乘后输入到语言模型中,最终训练得到汉字的输出结果. 模型整体实现了语音到汉字的端到端语音识别,在训练过程中对模型统一优化使用交叉熵作为损失函数,为了提高模型的泛化能力和学习能力采用标签平滑进行处理. 下面将对声学模型、语音帧分解模型和语言模型三个主要模块进行详细介绍.

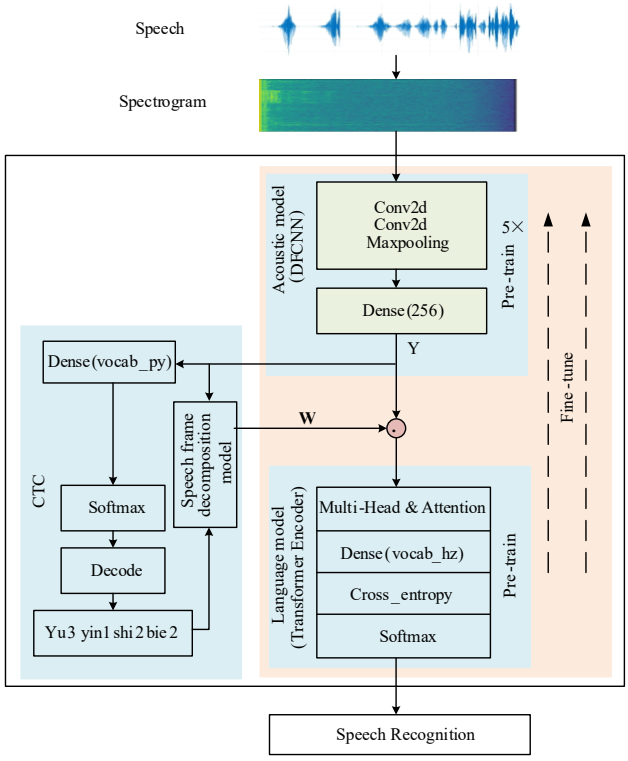

图1　融合语言模型的端到端框架

3.2　声学模型

在声学模型设计上采用深度全卷积神经网络(DFCNN)加上联结时序分类算法(CTC)的框架,通过堆叠多个卷积层直接对表示整句语音信号的语谱图进行建模,更好地表达了语音的长时相关性. 首先,从输入端来看,DFCNN摒弃了依赖人工经验设计的传统语音特征提取方法,直接利用从音频信号中提取出的语谱图作为输入保留了更多的原始语音信息,相比其他以传统语音特征

作为输入的语音识别框架具有天然的优势. 其次,从模型结构来看,DFCNN与传统语音识别中的CNN做法不同,它借鉴了图像识别中效果较好的网络配置,每个卷积层使用3×3的小卷积核,并在多个卷积层之后再加上池化层,大大增强了CNN的表达能力. 与此同时,通过累积较多的卷积池化层对,DFCNN可以看到远程的历史和未来信息,保证了DFCNN可以出色地表达语音的长时相关性,相比RNN网络结构在鲁棒性上更具优势. 模型的输出端采用CTC损失函数实现输入和输出的不等长对齐. 模型的具体结构如图2所示.

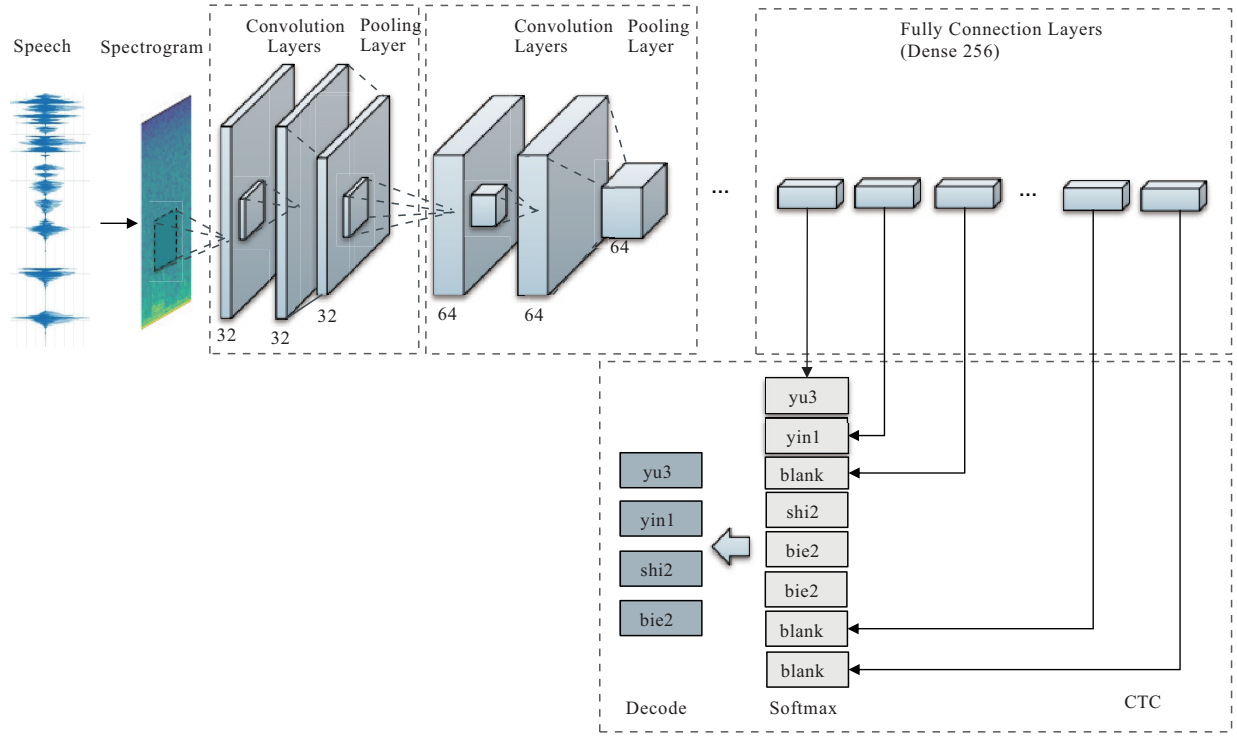

图2 DFCNN+CTC模型结构

模型整体实现语音文件到声学建模单元(有调音节如"da4 jia1 hao3"等)的转换过程,原始输入为一个语音片段,然后经过DFCNN的前端处理先对语音进行分帧加窗等预处理,设n时刻语音采样值为$x'(n)$,为预加重系数,$0.9<\alpha<1.0$,经过预加重处理后的结果如下式所示:

$$x(n)=x'(n)-\alpha x'(n-1) \tag{4}$$

然后对每帧进行傅里叶变换得到语音的语谱图X,直接将时间和频率作为图像的两个维度,对时频图进行卷积池化等特征提取操作,通过较多的卷积层和池化层的积累,DFCNN能看到足够长的历史和未来信息从而实现对整句语音的建模. 在模型参数配置上设置了10层卷积和5层池化,卷积核全部采用3×3的小卷积,不同层设置了不同的卷积核个数分别为32、64、128、128、128,在两次卷积之后再进行最大池化操作,前三层中pool_size均为2,后两层pool_size为1,池化的过程不使用激活函数. 为了增加网络的稳定性,在每次卷积操作之后都进行了批量归一化(Batch normalization, BN)操作,同时为了减少模型过拟合的风险,在网络中周期性地插入Dropout层. 模型在输出端采用CTC作为损失函数,自动实现语音帧和声学建模单元的自动对齐,以实现整个模型的端到端训练.

3.3 语音帧分解模型

声学模型与语言模型融合的难点在于声学模型和语言模型的优化目标不一致,声学模型需要CTC规则在训练过程实现语音帧和音节的自动对齐,在解码过程中实现语音帧和音节的转换,而语言模型需要交叉熵函数来计算损失. 所以需要一个额外的模型来完成两个模型优化目标的统一,也就是本文中的语音帧分解模型,通过该分解模型不需要CTC就能完成语音帧到音节的转换,从而使得整个模型有统一的损失函数.

语音帧分解模型的主要工作是计算一个权重矩阵W,其分量W_{ij}代表第j列音帧是否属于第i个音节序列,如果等于1表示第j列音帧对应第i个音节序列;如果有连续k列音帧同时对应第i个音节序列,如果有包括第j列在内的连续k列音帧同时对应第i个音节序列,则$W_{ij}=1/k$. 图3给出了算法的简单示意图,下面详细说明算法的具体过程.

语音经过特征提取之后进入卷积神经网络之前的输入是一维度为(batch_size, len_wav, len_feature, in_channels)的张量,batch_size表示一次性输入批处理语

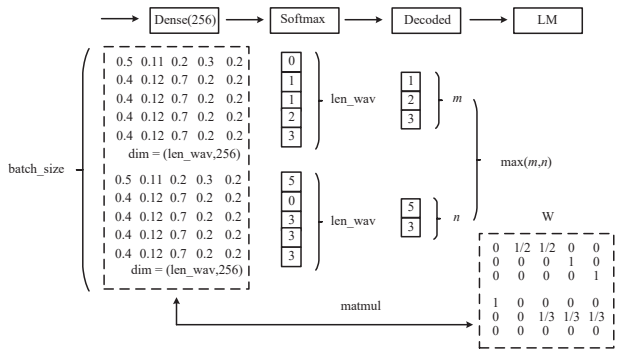

图3　权重矩阵计算过程示意图

音文件的个数，len_wav 表示最长语音文件的帧数，len_feature 表示语音文件经特征提取后每帧的特征维度，本文取值为 200，它与 len_wav 共同张成了所谓的“语谱图”，in_channels 为通道数，本文取值 1. 经过多层卷积池化及全连接层之后输出维度为（batch_size，len_wav，out_feature），其中 out_feature 表示单帧对应的输出特征维度，本文取值为 256，因此本层记为 Dense(256). 接下来经过 Dense(vocab) 计算之后取 Softmax 得到每一帧所对应音节的 one-hot 向量，其中 vocab 为音节集合的大小. 再经过 CTC 解码去重去空白得到最终输出的音节序列.

权重矩阵是用来刻画输出的拼音序列与输入每一帧的对应关系. 为方便讨论，后面不再考虑批处理和通道两个维度. 设 $\boldsymbol{L}\in\boldsymbol{R}^{\text{len\_wav}\times\text{vocab}}$ 为 Softmax 得到的 one-hot 向量排列而成的矩阵；$S'\in Z^{\text{len\_wav}}$ 为最终输出的音节序列，$s_i\in\{0,1,\cdots,\text{vocab}\}$ 表示第 i 帧对应的音节标号，当 $s_i=0$ 时表示音节序列结束，无对应音节，将无对应音节的 s_i 删除，可得最终输出的音节序列 $S\in Z^{\text{len\_yinjie}}$，len_yinjie 表示音节长度；$\boldsymbol{W}\in\boldsymbol{R}^{\text{len\_yinjie}\times\text{len\_wav}}$ 为语音帧对应输出音节的权重矩阵，$w_{ij}=0$ 表示音节 s_i 与语音帧 l_j 不对应，$w_{ij}=1$ 表示音节 s_i 恰好与语音帧 l_j 对应，若包括 l_j 在内的连续 k 个语音帧与 s_i 对应，则 $w_{ij}=1/k$. 为求权重矩阵 $\boldsymbol{W}$，首先将其初始化为 0 矩阵，之后从 s_1 到 $s_{\text{len\_yinjie}}$ 开始遍历 L 的每一行，记录下其在 L 中的位置，最后再进行归一化处理，详细算法由下面的伪代码给出.

权重矩阵 $\boldsymbol{W}\in\boldsymbol{R}^{\text{len\_yinjie}\times\text{len\_wav}}$ 代表语音帧与拼音序列直接的对应关系，将其与 Dense(256) 层的输出矩阵 $\boldsymbol{Y}\in\boldsymbol{R}^{\text{len\_wav}\times 256}$ 做矩阵相乘，可得

$$\boldsymbol{Y}'=\boldsymbol{W}\,\boldsymbol{Y} \tag{5}$$

即有 $\boldsymbol{Y}'\in\boldsymbol{R}^{\text{len\_wav}\times 256}$，并将其作为下一模块语言模型的输入. 由于 $\boldsymbol{Y}'$ 和 $\boldsymbol{Y}$ 的关系可由式(5)给出，因此在训练过程中，语言模型的损失函数可以直接反传回声学模型，从而实现端到端的学习过程.

3.4　语言模型

Transformer 是由 Google 公司提出，可解决机器翻译领域不定长序列映射问题的语言模型[21]. 该模型采用

算法 1　语音帧对应输出音节的权重矩阵计算

输入：语音帧矩阵 $\boldsymbol{L}\in\boldsymbol{R}^{\text{len\_wav}\times\text{vocal}}$，音节序列 $S\in Z^{\text{len\_yinjie}}$

输出：语音帧对应输出音节的权重矩阵 $\boldsymbol{W}$

```
FUNCTION SyllableWeightCal(L, S)
    W ← 0
    pos ← 1
    FOR i = 1 to len_wav DO
        cont ← True
        begin ← False
        WHILE cont AND pos ≤ len_wav
            IF S(i) = arg max_j L(pos, j)
                begin ← True
                W(i, pos) ← 1
            ELSE IF begin THEN
                cont ← False
            END IF
            pos ← pos + 1
        END WHILE
    END FOR
    W ← 按行归一化 W
    RETURN W
END FUNCTION
```

自注意力（Self-attention）机制的编码解码结构. 本文借鉴 Transformer 网络结构实现了从声学模型输出的音节序列到文字的解码过程. 相比 N-gram 模型，Transformer 网络更容易捕获句子中长距离的相互依赖特性，能充分利用语境信息，在音字转换中发挥更大的优势. Transformer 模型使用多头自注意力机制，具体描述如式(6)所示：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Soft max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{6}$$

其中 $\boldsymbol{Q}$ 表示查询向量，$\boldsymbol{K}$ 表示键向量，$\boldsymbol{V}$ 表示值向量，d_k 表示键向量的维度. 多头自注意力机制和传统注意力机制相比能够捕捉音节序列自身词与词之间的依赖关系，而与循环神经网络相比，具有更好的计算并行性，适合语言模型建模任务.

考虑音节序列与汉字是一一对应关系，不涉及到序列长度不相等的关系，所以 Decoder 端并不适合拼音转汉字这种定长序列的处理，因此本文只选择 Transformer 的 Encoder 结构并对其进行适当调整，即在输出部分增加一个全连接层和 Softmax 层，同时为了提升模型训练效果使用了标签平滑处理. 图 4 为本文所设计的语言模型结构.

语言模型整体实现音节序列到文字序列的解码过程，建模单元为汉字. 在预训练的过程中输入端为带音调的汉语音节序列，经过词嵌入层 Embedding 转化为对

应的词向量,词向量维度为256,并加入位置编码. 模型由多个相同的编码模块组成,每一个模块由多头自注意力Multi-Head Attention和一个全连接的前馈神经网络Feed-forward组成,此外每层采用残差之后对该层进行归一化(Layer-Normalization). 经过数据编码层之后先经过多头自注意力模块得到一个加权之后的特征向量,将其送到下一个前馈神经网络模块作为输入,这个全连接有两层,第一层的激活函数是Relu,第二层是一个线性激活层. 至此数据在每个模块中计算完毕,其输出作为下一个模块的输入进行相同的计算. 本文中多头注意力参数num_heads=8,对编码器使用6层自注意力模块进行堆叠,为了更好地学习网络对标签进行了标签平滑处理. 输出端采用Cross-entropy作为损失函数,以实现对整个网络的训练优化.

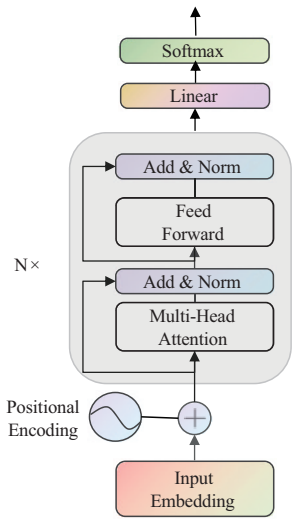

图4 基于Transformer编码器的语言模型

4 实验结果与分析

4.1 实验设计

4.1.1 实验数据与评价指标

本实验所用数据是由北京希尔贝壳科技有限公司开源的AISHELL语音数据集[22]. 该数据集录音文本涉及金融、房产、智能家居、工业生产等11个领域,音频是由400名来自中国不同口音区域的发言人在安静室内环境中使用高保真麦克风(44.1kHz,16bit)进行录制的,其中音频降采样为16kHz. 经过专业语音校对人员转写标注并通过严格质量检验,此数据库文本正确率在95%以上.

该数据集语音数据总时长为178h,本实验中,语音数据集划分为训练集、验证集和测试集三部分,其中训练集约为163h(130836条),验证集为5h(3588条),测试集为10h(7176条).

本文对语言模型的训练除了使用AISHELL数据集之外,还使用了搜狗短文本数据集[23],共5万句短文本数据,整个数据集内容涉及到新闻、时政、历史及房产等多个领域,整个数据集划分为训练集、验证集和测试集,由于训练语音模型需要拼音作为模型输入,所以在模型训练前需要进行数据处理,本实验通过调用Python的PyPinyin库将汉字转换为对应的拼音,经处理后得到覆盖4413个常用汉字和1372个带音调的拼音的中文训练语料.

本文使用的算法评估指标为字错误率WER,一般情况下WER越低表示识别性能越好. 评估标准具体如下所述:

对于语音识别的预测结果,一般需要对此结果进行替换、删除、插入某些词使之和标签词序列完全相同,修改的总词数除以标签词序列的个数即为WER,其计算公式如下:

$$\mathrm{WER}=100\%\times\frac{S+D+I}{N} \tag{7}$$

其中,S是被替换的单词数,D是被删除的单词数,I是被新插入的单词数,N是真实正确的单词数.

4.1.2 实验平台

本文实验均在Ubantu16.04(64bit)系统上运行,主要使用Python编程语言,结合Anaconda环境管理和包管理工具来配置不同的项目环境,语音处理库主要使用Librosa、Scipy等. 在构建模型结构时使用了Tensorflow内部高度封装的API Keras,为了提高训练速度使用Tensorflow-GPU版本,服务器GPU配置为4台GXT1080,运行内存为32GB. 为了更好的调试网络使用了Tensorflow2.0版本它的主要优点在于引进了动态度Eager模式,不再是1.x版本的静态图,对于内部的张量计算可以直接操作类似于Numpy数组.

4.1.3 模型训练及优化

该实验首先对所有语音数据进行预处理操作,包括预加重、分帧(帧长25ms、帧移10ms)及加窗(汉明窗),通过快速傅里叶变换提取语音语谱图作为输入特征,共200维. 在训练阶段选取适应性动量估计算法(Adaptive moment estimation, Adam)作为模型的优化器,该算法不仅能够对不同参数计算适应性学习率,而且能够加速网络收敛速度[24];在每层卷积层之后添加批量归一化(BN)对网络中的权重进行自适应调整,以此提高网络的训练速度和泛化能力[25];在池化层之后使用Dropout[26]以此有效地降低网络的过拟合风险,初始学习率设置为1×10^{-3};在微调阶段,以随机梯度下降算法(Stochastic Gradient Descent, SGD)作为模型的优化器,通过设置更小的学习率使得网络在后期优化更为稳定,微调学习率设置为1×10^{-5}.

4.1.4 实验对比模型

为了验证本文所提出的融合语言模型进行中文语

音识别的整体框架效果以及所提出的端到端训练算法的作用,分别设计了两组消融实验,第一组是不包含语言模型的从语音到汉字的DFCNN端到端模型,第二组结构上与本文提出方法基本相同,由从语音到拼音的DFCNN模型和从拼音到汉字的语言模型构成,但是两个模型需要分别进行训练,不包含整体的端到端训练过程. 这两组实验对比模型分别记为DFCNN和DFCNN+Transformer Encoder.

同时本文也和当前五种主流算法进行了对比,即BLSTM-DNN-CTC模型[27]、CNN-LSTM-DNN-CTC模型[14]、DCNN-DNN-CTC模型[9]、ResNet-BLSTM模型[28]和Deep Speech2模型[15]. BLSTM-DNN-CTC通过使用具有CTC输出层的深层双向LSTM网络处理频谱图来解决语音识别问题. CNN-LSTM-DNN-CTC将CNN、LSTM和DNN组合起来以利用三者在建模能力上的互补关系,CNN擅长减少频率变化,LSTM擅长执行时间建模,DNN适合用于将特征表示映射到更可分离的空间. DCNN-DNN-CTC通过将分层的CNN与CTC直接结合来实现一种序列标记的端到端语音识别框架. ResNet-BLSTM在CNN中引入残差模块和并行的卷积层并接入BLSTM层,前者能够提取语谱图中的局部区域特征,由后者对特征进行上下文建模. Deep Speech2的架构是一个递归神经网络,具有一个或多个卷积输入层和多个递归层. 本文方法及7个对比模型的结构参数见表1. 模型DFCNN、DFCNN+Transformer Encoder、DCNN-DNN-CTC、ResNet-BLSTM及我们提出的ECSRILM模型输入均为语谱图. BLSTM-DNN-CTC采用梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient,MFCC)特征参数为26维,输入为前后9帧加当前帧一共494维参数;模型CNN-LSTM-DNN-CTC和Deep Speech2输入特征参数为包含一阶、二阶差分共120维梅尔标度滤波器组特征(Mel-scale Filter Bank,FBank),其特征提取方法类似于MFCC,是将MFCC最后一步的离散余弦变换去掉而得到的语音特征,保留了更多的原始语音信息. 模型DFCNN、DFCNN+Transformer Encoder、Deep Speech2以及本文方法是作者采用AISHELL数据集进行训练,并在测试集上进行性能测试得到的实验结果;而余下模型的结果为文献[28]中基于相同的AISHELL数据集进行训练和测试得到结果.

表1　不同模型结构参数

模型	网络结构参数	声学建模单元
DFCNN	5×2D_CNN+2×FNN	汉字
DFCNN+Transformer Encoder	5×2D_CNN+2×FNN+Transformer Encoder	带音调音节
BLSTM-DNN-CTC[27]	5×DNN+3×BLSTM+2×FNN	汉字
CNN-LSTM-DNN-CTC[14]	3×CNN+3×LSTM+4×FNN	汉字
DCNN-DNN-CTC[9]	10×CNN_maxout+3×FNN_maxout	汉字
ResNet-BLSTM[28]	8×CNN(4×Res)+2×BLSTM	汉字
Deep Speech2[15]	3×1D_CNN+3×GRU+FNN	汉字
ECSRILM	5×2D_CNN+2×FNN+Transformer Encoder	汉字

4.2　实验结果

在本节中分别给出了消融实验的纵向实验对比结果和与现有主流方法的横向实验对比结果,采用字错误率WER作为性能评价指标,WER的值越小表示识别性能越好.

4.2.1　纵向对比实验

消融实验的结果如表2所示. DFCNN是直接对汉字进行建模,缺少语言模型,其WER达到15.06%,结果中出现了很多多音字替换错误,本模型由于引入了语言模型进行辅助解码,WER为11.88%,与DFCNN的WER相比下降幅度为21%,在一定程度上对同音异义字进行了区分;DFCNN+Transformer Encoder采用字级别的浅层融合方法,虽然利用了外部语言模型,但只是在最后解码阶段才加入的,并没有考虑在模型训练阶段语言模型对声学模型的影响,语言模型并没有对声学模型的错误结果进行纠正也没有参与到声学模型的训练过程中;同时由于声学模型和语言模型是分开训练的,声学模型产生的误差直接影响了语言模型,而且使用不同的优化函数在训练过程中也不便于统一优化增加了整个系统构建的工作量,因此WER反而比DFCNN更高,达到了20.72%. 本模型将语言模型和声学模型深层融合并进行联合训练,使得语言模型在训练和预测阶段都对声学模型进行了纠正,同时模型的统一优化也减少了中间的误差损失,与DFCNN+Transformer Encoder的WER相比下降幅度为43%. 实验结果表明本文所提出的融合语言模型的识别框架有效地提高了中文语音识别的精度,其中端到端的训练算法是关键.

表2　纵向对比实验结果

模型结构	字错误率(%)
DFCNN	15.06
DFCNN+Transformer Encoder	20.72
ECSRILM	11.88

4.2.2　横向对比实验

本文方法和当前五种主流算法的实验对比结果如表3所示,其中Deep Speech2的结果通过开源复现代码在AISHELL数据集上训练得到,其他四种模型的实验结果为文献[28]中给出的结果. 从表3中可以看出,本文提出的ECSRILM模型的字错误率远低于其他4种对

比模型,比其中最好的DeepSpeech2模型降低了28%.其主要原因是ECSRILM模型采用二维卷积直接对语音的频谱图进行特征提取,在一定程度上缓解了传统声学特征提取方法中过分依赖经验而造成特征信息部分丢失的情况.采用多个小卷积核来代替大卷积核增加了模型的非线性区分能力,通过累积非常深的卷积池化操作使得模型能学习到更多的历史信息,更加充分地表达语音的长时相关性.同时由于ECSRILM模型在解码过程中增加了语言模型,通过引入先验的语言学信息来纠正中文语言的同音异义字的替换错误从而提高了模型的识别性能.

表3 横向对比实验结果

模型	字错误率(%)
BLSTM-DNN-CTC[23]	26.27
CNN-LSTM-DNN-CTC[11]	24.55
DCNN-DNN-CTC[9]	23.36
ResNet-BLSTM[24]	20.84
Deep Speech2[12]	16.60
ECSRILM	11.88

5 结论

本文针对CTC端到端模型缺少外部语言模型的指导容易出现同音字或近音字错误等问题,提出了一种融合语言模型的端到端中文语音识别方法ECSRILM.其中声学模型采用DFCNN框架,语言模型则是基于Transformer结构中的编码器部分设计而成.针对声学模型和语言模型无法联合优化的难点,我们设计了语音帧分解模型,实现了声学模型的语音帧输出和音节模型的匹配,并作为语言模型的输入.该模型能够实现损失函数由语言模型反传至声学模型,进而实现整个模型的端到端训练过程.

为验证本文方法的有效性,设计了两组消融实验进行纵向比较,并与其他五种主流方法进行了横向对比.实验结果表明,所提出的融合语言模型的识别框架有效地提高了中文语音识别的精度,其中端到端的训练过程是算法的关键环节.该方法也全面优于现有的五种主流方法,表明在一定程度上纠正了基于CTC语音识别系统产生的多音字替换错误,整体上提高了识别正确率.

参考文献

[1] 杨明浩,高廷丽,陶建华,等.对话意图及语音识别错误对交互体验的影响[J].软件学报,2016,27(S2):69－75.
Yang MH, Gao TL, Tao JH, et al. Error analysis of intention classification and speech recognition in human-computer dialog [J]. Journal of Software, 2016, 27(S2): 69－75. (in Chinese)

[2] Rodriguez E, RuiZ B, Garcia-Crespo A, et al. Speech/speaker recognition using a HMM/GMM hybrid model [A]. International Conference on Audio-and Video-Based Biometric Person Authentication [C]. Berlin, Heidelberg: Springer, 1997. 227－234.

[3] Mohamed AR, Sainath TN, Dahl G, et al. Deep belief networks using discriminative features for phone recognition [A]. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing [C]. Prague, Czech Republic: IEEE, 2011. 5060－5063.

[4] Yu D, Deng L. Deep learning and its applications to signal and information processing [J]. IEEE Signal Processing Magazine, 2011, 28(1): 145－154.

[5] Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks [A]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [C]. Vancouver, Canada: IEEE, 2013. 6645－6649.

[6] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [A]. The 15th Annual Conference of the International Speech Communication Association [C]. Singapore: ISCA, 2014.338－342.

[7] Abdel-Hamid O, Mohamed AR, Jiang H, et al. Convolutional neural networks for speech recognition [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(10): 1533－1545.

[8] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [A]. International Conference on Machine Learning, ICML 2006 [C]. Pittsburgh, PA: ACM, 2006. 369－376.

[9] Zhang Y, Pezeshki M, Brakel P, et al. Towards end-to-end speech recognition with deep convolutional neural networks [A]. The 17th Annual Conference of the International Speech Communication Association [C]. San Francisco, CA: ISCA, 2016. 410－414.

[10] Yang XD, Wang WZ, Yang HW, et al. Simple data augmented transformer end-to-end Tibetan speech recognition [A]. IEEE 3rd International Conference on Information Communication and Signal Processing [C]. NY: IEEE, 2020. 148－152.

[11] Chang HJ, Liu AH, Lee HY, et al. End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training [A]. IEEE Spoken Language Technology Workshop [C]. NY: IEEE, 2021. 186－193.

[12] Fan CH, Yi JY, Tao JH, et al. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 198 − 209.

[13] Graves A, Jürgen S. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5-6): 602 − 610.

[14] Sainath TN, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks [A]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [C]. NY: IEEE, 2015. 4580 − 4584.

[15] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: end-to-end speech recognition in English and Mandarin [A]. International Conference on Machine Learning 2016 [C]. NY: ACM, 2016. 173 − 182.

[16] 王海坤,潘嘉,刘聪.语音识别技术的研究进展与展望[J].电信科学报, 2018, 2: 1 − 11.
Wang HK, Pan J, Liu C. Research development and forecast of automatic speech recognition technologies [J]. Telecommunications Science, 2018, 2: 1 − 11. (in Chinese)

[17] Kannan A, Wu YH, Nguyen P, et al. An analysis of incorporating an external language model into a sequence-to-sequence model [A]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing [C]. Calgary, Canada: IEEE, 2017. 5824−5828.

[18] Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation [OL]. http://arxiv.org/abs/1503.03535, 2015.

[19] Anuroop S, Heewoo J, Sanjeev S, et al. Cold fusion: Training seq2seq models together with language models [A]. The 19th Annual Conference of the International Speech Communication Association [C]. Hyderabad, India: ISCA, 2018. 387 − 391.

[20] Toshniwal S, Kannan A, Chiu CC, et al. A comparison of techniques for language model integration in encoder-decoder speech recognition [A]. IEEE Workshop on Spoken Language Technology [C]. Athens, Greece: IEEE, 2018. 369 − 375.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [A]. Advances in Neural Information Processing Systems[C]. Long Beach, CA: MIT Press, 2017. 5998−6008.

[22] Bu H, Du J, Na X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline [A]. The 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA) [C]. Seoul, South Korea: IEEE, 2017. 58 − 62.

[23] Wang CH, Zhang M, Ma SP, et al. Automatic online news issue construction in Web environment [A]. The 17th International World Wide Web Conference [C]. Beijing, China: ACM, 2008. 457 − 466.

[24] Kingma D, Ba J. Adam: a method for stochastic optimization [A]. IEEE 17th International Conference on Computational Science and Engineering (CSE) [C]. Chengdu, China: IEEE, 2014. 563 − 568.

[25] Sergey I, Christian S. Batch normalization: accelerating deep network training by reducing internal covariate shift [A]. International Conference on Machine Learning 2015 [C]. Lille France: ACM, 2015. 448 − 456.

[26] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929 − 1958.

[27] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [A]. International Conference on Machine Learning[C]. Beijing, China: JMLR, 2014, 32(2): 1764 − 1772.

[28] 胡章芳,徐轩,付亚芹,等.基于ResNet-BLSTM的端到端语音识别[J].计算机工程与应用, 2020, 56(18): 124−130.
Hu ZF, Xu X, Fu YQ, et al. End to end speech recognition based on ResNet-BLSTM[J]. Computer Engineering and Applications, 2020, 56(18): 124 − 130. (in Chinese)

作者简介

吕坤儒　男,1993年生,河南安阳人.吉林大学软件学院硕士研究生.研究方向:机器学习.
E-mail:892539843@qq.com

时小虎(通信作者)　男,1974年生,河北玉田人.吉林大学计算机科学与技术学院教授.研究方向:机器学习.
E-mail:shixh@jlu.edu.cn