计算机科学
COMPUTER SCIENCE

融合音字特征转换的非自回归Transformer中文语音识别

滕思航, 王烈, 李雅

引用本文

滕思航, 王烈, 李雅.融合音字特征转换的非自回归Transformer中文语音识别[J]. 计算机科学, 2023, 50(8): 111-117.

TENG Sihang, WANG Lie, LI Ya. Non-autoregressive Transformer Chinese Speech Recognition Incorporating Pronunciation- Character Representation Conversion [J]. Computer Science, 2023, 50(8): 111-117.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600144

融合音字特征转换的非自回归 Transformer 中文语音识别

滕思航　王　烈　李　雅
广西大学计算机与电子信息学院　南宁 530004
(2013391124@st.gxu.edu.cn)

摘　要　基于自注意力机制的 Transformer 模型在语音识别任务中展现出了强大的模型性能,其中非自回归 Transformer 自动语音识别模型与自回归模型相比解码速度更快,然而语音识别速度的提升却造成了准确度的大幅降低。为提升非自回归 Transformer 语音识别模型的识别准确度,首先引入基于连续时间分类(Connectionist Temporal Classification,CTC)的帧信息合并,在帧宽范围内对语音高维表示向量进行融合,改善非自回归 Transformer decoder 输入序列的特征信息不完整问题;其次对模型输出进行音字特征转换,在 decoder 的输出读音特征中融合上下文信息,然后转换为包含更多字符特征的输出,从而改善模型同音不同字的识别错误问题。在中文语音数据集 AISHELL-1 上的实验结果显示,所提模型实现了实时性因子(Real Time Factor,RTF)0.0028 的识别速度与字符错误率(Character Error Rate,CER)8.3%的识别精度,在众多主流中文语音识别算法中展现出较强的竞争力。

关键词:语音识别;Transformer;非自回归;自注意力机制;特征转换
中图法分类号　TP391

Non-autoregressive Transformer Chinese Speech Recognition Incorporating Pronunciation-Character Representation Conversion

TENG Sihang,WANG Lie and LI Ya
School of Computer,Electronics and Information,Guangxi University,Nanning 530004,China

Abstract　The Transformer based on self-attention mechanism shows powerful model performance in speech recognition tasks, where the non-autoregressive Transformer automatic speech recognition model has a faster decoding speed compared with the autoregressive model. However, the increase in speech recognition speed causes a larger decrease in accuracy. To improve the accuracy of the non-autoregressive Transformer speech recognition model, the frame information merging based on connectionist temporal classification(CTC) is introduced firstly, which fuses the speech high-dimensional representation in the frame width range to improve the problem of incomplete feature information in the non-autoregressive Transformer decoder input sequences. Secondly, pronunciation-character representation conversion is performed on the model output, and the pronunciation representation is converted into an output containing more character features by fusing contextual information on the pronunciation features of the decoder output, thus improving the recognition error problem of the model with different characters in the same pronunciation. Experiments on the Chinese speech dataset AISHELL-1 show that the proposed model achieves a recognition speed of real time factor(RTF) 0.0028 and recognition accuracy of 8.3% character error rate(CER), demonstrating strong competitiveness among many mainstream Chinese speech recognition algorithms.

Keywords　Speech recognition, Transformer, Non-autoregressive, Self-attention mechanism, Representation conversion

1　引言

近年来,智能设备呈现出蓬勃发展的趋势,已经成为人们在日常生产生活中不可或缺的一部分。语音交流具备快速、准确的特点,是人与人之间的一种重要沟通方式,利用语音在人与智能设备之间进行交互也因此成为一项重要的研究内容[1]。传统的自动语音识别(Automatic Speech Recognition,ASR)系统如基于隐马尔可夫的深度神经网络[2](Deep Neural Network-Hidden Markov Models,DNN-HMM)模型由声学模型(Acoustic Model,AM)和语言模型(Language Model,

到稿日期:2022-06-16　返修日期:2023-04-23
基金项目:广西科技重大专项(桂科 AA21077007-1)
This work was supported by the Science and Technology Key Projects of Guangxi Province(AA21077007-1).
通信作者:王烈(lwang@gxu.edu.cn)

LM)构成,已经被成功应用于 ASR 任务[3]。然而,具有混合模型结构的 DNN-HMM 需要预先对高斯混合模型进行单独训练,发音字典也需要语言学专家的知识才能建立,AM 和 LM 模型的单独训练使得模型难以进行联合优化。上述难点导致模型识别效果不尽人意,阻碍了 ASR 的进一步发展。

混合模型的复杂性使得研究人员开始关注端到端的 ASR 模型[4-5]。相较于混合模型,端到端的系统只由一个模型构成,并且模型的构建不需要专家知识,模型的训练难度和复杂度得到了简化,ASR 的性能因此得到了较大提升。最近,基于注意力机制的深度神经网络模型[6-8]展现出了优秀的性能,其中具有代表性的是采用 encoder-decoder 结构的基于注意力机制的 Transformer 神经网络模型。Transformer 首先成功应用于机器翻译任务中[9-10],研究人员受其启发将其拓展到 ASR 任务后也取得了不错的成绩[11]。这种完全利用注意力机制的模型可以获取长序列的时序信息,相比 LSTM[12](Long Short-Term Memory)和 GRU[13](Gated Recurrent Unit)等循环神经网络结构(Recurrent Neural Network,RNN),Transformer 的并行化计算可以充分利用先进的显卡等硬件的计算能力,实现更高效的模型优化和更快的识别计算。然而,语音识别 Transformer 在每一次计算时只能得到一个输出标签,这种一步步获取完整结果的解码方式被称为自回归式[14]的解码。因此当目标序列过长时,计算时间也会相应增加,识别速度会受到极大影响[15]。

为了进一步加快 ASR 的识别速度,研究人员对解码的方式进行了改进,将原本自回归式的解码调整为非自回归式解码[16-18],ASR 在比以前短得多的时间内就可以获取完整的输出结果,识别速度有了显著的提高。然而,速度的提升是以牺牲部分识别准确率为代价换取而来的,非自回归 Transformer(Non-Autoregressive Transformer,NAT)的准确率与自回归 Transformer(Autoregressive Transformer,AT)之间有很大的差距,这是我们不愿意看到的结果。

本文研究是基于上述问题开展的。本文的研究目的主要是利用 NAT 较快的解码速度,并尽可能提升 NAT 模型的识别准确率,达到与 AT 模型相近甚至更好的性能。通过对前序 NAT 模型的识别结果出错类型进行分析,本文提出了一种音字特征转换的深度上下文语境表征模型,在 decoder 输出端捕捉读音特征的句法信息,动态生成字特征,将传统模型输出的读音特征转换为唯一的字特征,从而提升 NAT 模型的识别准确性。

2 相关工作

本节对语音识别 Transformer 模型的解码方式进行简要介绍。

2.1 自回归 Transformer ASR

传统 Transformer 自动语音识别模型采用了自回归式的解码方式,其过程如图 1 所示。在解码阶段计算当前时刻 t 的输出时,模型需要 encoder 的语音高维表示向量和历史输出共同作为输入,即:

$$\hat{y}_t=\text{argmax}(P(y_t|y_{<t},x)) \tag{1}$$

其中,$\hat{y}_t$ 为在当前时刻 t 的输出结果。在预测第一个字符时由于无初始值,此时只能利用句头标签〈SOS〉作为 $y_{<t}$ 的输入值,当预测结果 $\hat{y}_t$ 出现句子结束标签〈EOS〉时结束。可以看出,AT 的一次计算只能产生一个输出字符标签,而一句语音所对应的转录文本往往包含一定数量的字符。当字符数量较多时,AT 的计算时间将会成倍增加,导致实时性无法达到令人满意的效果。计算过程中还存在着另一个问题,在训练阶段,$y_{<t}$ 的输入值是训练集的正确对应转录字符,然而在预测阶段没有正确转录文本,只能利用已经产生的历史输出结果作为输入值,这就造成了在训练和预测阶段的不匹配问题[19],因此出现的错误累积问题也会影响识别结果的准确率。

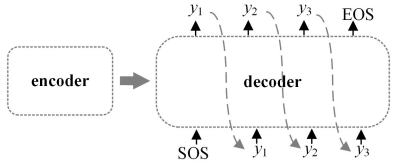

图 1 自回归 Transformer 解码

Fig. 1 Decoding of AT

2.2 非自回归 Transformer ASR

为了进一步减少自动语音识别模型解码的计算时间,研究人员提出了非自回归式解码的 Transformer 模型,NAT 的计算过程如图 2 所示,与 AT 对历史信息产生依赖不同,NAT 在解码时只需经过一次迭代计算就能获取完整的转录结果 $\hat{Y}$,解码时间得到了成倍的缩减。计算式如下:

$$\hat{Y}=\text{argmax}\ P(Y|X) \tag{2}$$

从式(2)可以看出,完整转录结果只是输入语音序列的函数。然而,为实现目标序列的一次性生成,NAT 的解码过程必须假设字符之间条件独立,每个标签的产生不依赖于其他字符,而字符之间的关系可以被认为是一种语法信息,独立性假设会使得模型忽略字词之间的依赖关系,导致识别结果准确性降低。

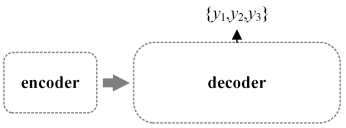

图 2 非自回归 Transformer 解码

Fig. 2 Decoding of NAT

3 本文方法

本节对所提出的基于音字特征转换的非自回归 Transformer 端到端中文语音识别算法进行介绍,模型的整体结构如图 3 所示。模型主要由 encoder、帧合并网络、decoder 和音字特征转换网络构成。语音信号首先经过语音处理层转换为语音特征序列,encoder 将语音特征序列提取为高维的语音表示向量,这些向量经过帧合并网络的特征融合后即可输入 decoder,最终经过音字特征转换后生成最终的识别字符标签序列。下面对各个模块进行详细分析。

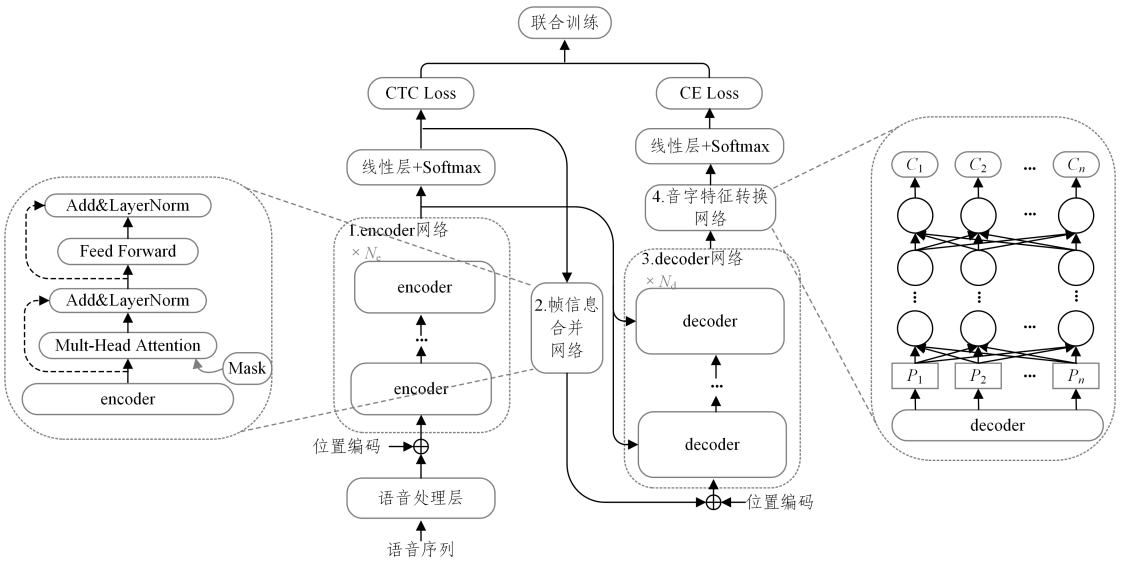

图 3 本文提出的模型框图

Fig. 3 Architecture of the proposed model

3.1 encoder 网络

将数据输入模型进行处理前，需要将原始语音波形信号处理为 F-bank[20] 声学特征，经过由两层二维卷积神经网络构成的语音处理层后，特征序列将在时间维度完成下采样，降低序列长度，向量维度也转换为与模型维度相等。接着通过对特征序列添加三角函数位置编码后即可输入 encoder。三角函数位置编码的计算式为：

$$pe_{i,2j}=\sin\left(\frac{i}{10\,000^{2j/D_m}}\right)$$
$$pe_{i,2j+1}=\cos\left(\frac{i}{10\,000^{2j/D_m}}\right) \tag{3}$$

其中，i 表示帧的序号，j 表示帧向量的元素序号，D_m 表示模型维度。由于 encoder 中的自注意力机制网络无法捕获序列的位置信息，位置编码的加入可以使模型学习到不同位置之间的相对关系。

encoder 由多个注意力层堆叠而成，注意力层又由多头自注意力子层以及前馈层构成，其中的查询、键和值向量为同一向量，因此可以实现自制注意力计算，并行化地提取语音的特征信息，encoder 的输出为语音高维表示序列。

3.2 CTC 帧信息合并网络

由于非自回归式解码无法使用历史输出信息，因此模型采用了基于 CTC 的帧信息合并网络来获取包含原始语音信息的 decoder 输入序列。

首先，利用 CTC[21] 对 encoder 进行训练，完成优化后的 encoder 提取出输入序列的特征，再经过全连接线性层与 Softmax 激活函数，生成二维的特征向量。本文将该二维特征向量称为帧信息向量，帧信息向量用于计算每一个语音高维表示向量的标签：有用标签{t}、噪声等其他无用标签{—}。当获得 encoder 输出向量标签后，将进行帧信息计算，帧信息计算过程如图 4 所示。其中帧宽范围为{—}后的第一个{t}到下一个{—}的前一个{t}，并以帧宽内概率值最大的帧作为中心帧，中心帧的数量将指示出语音信号中所包含的字符数量。帧宽与中心帧信息将用于下一步的帧信息合并。

可以看出，encoder 的训练目标为标签{t}或{-}，因此我们将训练集所有目标字符标签转变为标签{t}后作为 CTC 损失的训练目标。此时，模型同时存在两个训练任务以至于难以进行训练，为了解决此问题，模型最终采用联合优化[22] 的策略，利用 CTC 损失函数和交叉熵损失函数来对模型进行优化，计算式如下：

$$\mathcal{L}=\alpha\mathcal{L}_{\mathrm{CE}}+(1-\alpha)\mathcal{L}_{\mathrm{CTC}} \tag{4}$$

其中，α 是一个超参数权重，采用联合优化策略避免了模型无法训练的问题，并且可以加快模型的收敛速度，简化训练过程。

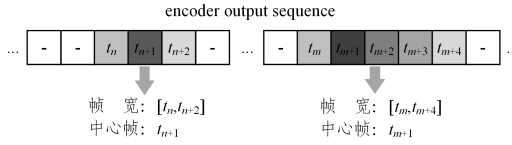

图 4 帧信息计算

Fig. 4 Frame information computation

帧信息合并网络由多头自注意力层构成，为了减少模型深度，我们采用单层结构。相比其他类型的神经网络，自注意力机制具有快速的并行化计算和较强的语言建模能力的特点。语音高维表示在每一个帧宽范围内经过合并器的融合后成为一个新的字符特征向量，图 3 中的 Mask 限定了帧信息的合并范围，新生成的向量按顺序组合后将作为 decoder 的输入序列。

3.3 decoder 网络

与 encoder 类似，decoder 同样通过堆叠注意力层实现语音特征的解码计算，其不同点在于自注意力子层与前馈层之间增加了一个交叉多头自注意力子层。decoder 以 encoder 输出的高维语音表示与表征合并网络输出的序列共同作为输入，通过注意力计算将编码后的特征向量进行解码，在本文模型中，将解码后的输出向量称为发音特征向量。

3.4 音字特征转换网络

对语音识别的错误结果进行分析后发现，大多产生替换错误的输出字符与正确转录字符的读音相近甚至相同，根据

该现象可以判断，模型能够正确提取语音信号的读音特征，但在计算输出字符标签概率时，静态的词表征模型无法捕捉字符的上下文特征信息，导致同音不同字的特征向量最终映射为同一个字符标签，产生错误的识别结果。为解决这一问题，我们观察人类的语言交流，发现当人们单独说出某个字符的读音时，往往难以准确判断对应的字符，而将其置于特定的语境中时即可准确判断出读音对应的字符。根据该原理，提出在模型的后端对输出向量进行音字特征转换后再生成最终结果，转换网络的作用正是将原本包含更多读音特征的输出转换为符合上下文语境的唯一字符特征输出，由此来改善同音不同字的错误识别问题，提高模型的识别准确率。

受预训练语言模型的启发[23-24]，本文引入两种不同的网络结构，即 Bi-LSTM 与深度 Bi-Transformer，将其作为提出的音字特征转换模型，通过实验比较两者的性能后获取较优的模型结构。音字特征转换的原理结构如图 3 所示，可以看出，特征转换网络中的每一个隐向量与其之前与之后的节点相连，双向连接的结构可以令隐向量具备上下文相关性，从而更好地提取字符的特征信息。假设 decoder 的输出包含 n 个向量$\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n\}$，则正向连接模型的输出条件概率为：

$$p(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n)=\prod_{k=1}^{n}p(\boldsymbol{P}_k\mid\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_{k-1}) \tag{5}$$

类似地，反向连接模型的输出条件概率为：

$$p(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n)=\prod_{k=1}^{n}p(\boldsymbol{P}_k\mid\boldsymbol{P}_{k+1},\boldsymbol{P}_{k+2},\cdots,P_n) \tag{6}$$

双向计算可以令每一个特征向量具有上下文相关性，并且多个双向连接层的堆叠可以提取不同维度的语法信息，融合了上下文特征的向量即可实现从读音特征到字特征的转换。下面详细介绍基于 Bi-LSTM 与基于深度 Bi-Transformer 的音字特征转换网络的原理。

3.4.1 Bi-LSTM

基于 Bi-LSTM 的音字特征转换模型如图 5 所示，每个节点都由 LSTM 单元构成。可以看出，输入向量 $\boldsymbol{P}_n$ 在双向连接的结构下，网络将计算出每一个特征向量的正向隐向量$\overrightarrow{h_n}$和反向隐向量$\overleftarrow{h_n}$，将其拼接成$[\overrightarrow{h_n},\overleftarrow{h_n}]$的形式后作为新的表征向量，因此生成的新表征向量融合了上文与下文的信息，从而可以更好地表示在当前语境下的字符特征。并且，随着层数的增加，每一层的隐向量将包含更深的含义[23]，假设输入为字符特征，则第一层的隐向量包含较多的句法特征，第二层的隐向量包含较多的语义信息。

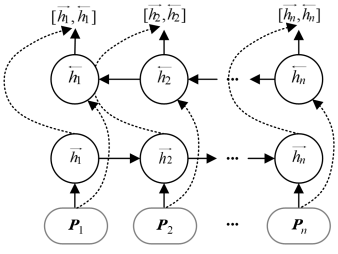

图 5 Bi-LSTM 模型

Fig. 5 Bi-LSTM model

根据以上描述，只需经过一层双向语言模型的计算，特征转换模型的输入读音特征即可转换为包含句法信息的读音特征，即字符特征。因此，考虑到要尽可能减小模型规模并加快训练和解码的速度，基于 Bi-LSTM 的特征转换模型采用单层结构，最终转换后的输出表征由输入读音特征向量与对应隐向量进行拼接后得到，与残差连接类似。

3.4.2 深度 Bi-Transformer

在最近的 NLP(Natural Language Processing)任务中，自注意力机制展现出了强大的语言建模能力和特征提取能力，因此本文引入自注意力网络作为音字特征转换模型的另一种方案。其中，自注意力块由多头注意力层和前馈层堆叠而成，多头注意力层可以对不同帧之间进行缩放点积注意力计算，并且每一个注意力头分别计算不同的注意力参数，计算式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{7}$$

其中，$\boldsymbol{Q}\in\mathbb{R}^{n\times d_q}$，$\boldsymbol{K}\in\mathbb{R}^{n\times d_k}$，$\boldsymbol{V}\in\mathbb{R}^{n\times d_v}$ 分别代表查询、键和值向量，n 和 d 分别是语音序列的长度和对应向量维度。为实现自注意力计算，采用与编码器类似的方法，$\boldsymbol{QKV}$ 为同一特征向量。缩放操作是为了限制计算值的大小并使模型训练更加稳定。

多头注意力的每一个头通过在不同的子空间进行注意力计算，从而获取多种不同的注意力信息，计算式为：

$$\begin{aligned}&MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_H)\boldsymbol{W}^O\\&head_h=Attention(\boldsymbol{Q}\boldsymbol{W}_h^Q,\boldsymbol{K}\boldsymbol{W}_h^K,\boldsymbol{V}\boldsymbol{W}_h^V)\end{aligned} \tag{8}$$

其中，$\boldsymbol{W}_h\in\mathbb{R}^{d\times d}$，$\boldsymbol{W}^O\in\mathbb{R}^{d_m\times d_m}$ 代表乘积矩阵，H 是多头数量，多头自注意力的计算过程如图 6 所示。

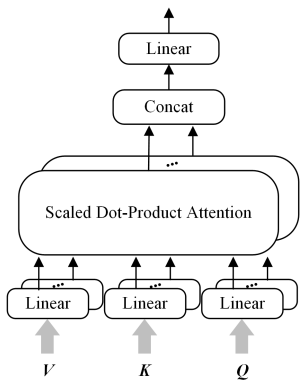

图 6 多头自注意力机制

Fig. 6 Multi-head self-attention mechanism

基于自注意力机制的音字特征转换网络与 Bi-LSTM 音字特征转换网络结构类似，通过堆叠多层自注意力层可以实现对特征的深度上下文信息提取。由式(7)可以看出，自注意力机制对序列进行了并行化的计算，并且每一个向量的注意力计算的范围囊括了上下文向量，使得每一个特征向量都包含了序列的上下文信息，从而动态生成字符特征。

4 实验

4.1 数据集

本文的所有实验都在公开中文语音数据集 AISHELL-1[25]上进行。

该数据集共包含 178h 的普通话语音信号，所有语音经过

44.1 kHz 的麦克风采样后再下采样为 16 kHz。AISHELL-1 中的数据分布如表 1 所列。其中，训练集对 340 名说话者录音，共包含 120098 个句子约 150 h 的时长；开发集对 40 名说话者录音，共包含 14326 个句子约 18 h 的时长；测试集一共 7176 个句子约 10 h 的时长。

表 1 AISHELL-1 数据分布

Table 1 Distribution of AISHELL-1

	句子	时长/h	说话者
训练集	120098	150	340
开发集	14326	18	40
测试集	7176	10	20

4.2 模型配置

字符标签的字典大小为 4233，由训练集中出现的所有字符、句子开头标签〈SOS〉、句子结尾标签〈EOS〉和未知符号标签〈UNK〉组成。模型训练时采用基于语音时长的采样，每批输入的时长总和为 150s，累积 8 批的梯度后再进行 1 次优化，使训练更稳定[26]。模型以 80 维梅尔尺度 F-bank 特征作为输入，每帧截取的窗长为 25 ms，帧移为 10 ms，并在时间和频率两个维度上进行语音数据增强[27]，提高模型泛化能力。语音信号在输入 encoder 前先经过语音处理层进行预处理，利用两层卷积核为 32 维、大小为 3、滑动为 2 的二维卷积层，可以得到 4 倍的帧数下降，序列维度也转换成与模型相同。在 encoder、帧合并器和 decoder 前都添加位置编码信息。所有实验的 encoder 和 decoder 都堆叠 6 层的注意力层，每个注意力层的头数为 8，前馈层的维度为 2048，模型维度为 512，激活函数为 GLU。

本文以 CTC 帧信息合并模型作为非自回归语音 Transformer 基线模型。合并器由 1 层多头注意力层构成，头数为 8，前馈网络维数为 2048，模型维数为 512，联合训练的超参数权重 α 设置为 0.4。对于音字特征转换模型，我们分别设置了在 3.2 节中介绍的两种结构进行实验比较。其中，基于 Bi-LSTM 结构的特征转换模型采用单层双向 LSTM 单元，每个单元的隐变量维度为 4096，并采用 0.1 的 dropout 和层标准化，输出的隐向量进行正反向拼接后与输入向量相加得到转换后的输出表征；基于 Bi-Transformer 的特征转换网络共包含 12 层多头自注意力块，模型维度为 768，每个多头自注意力模块头数为 12，前馈层的维度为 2048，并采用 0.1 的 dropout 和 GeLU 激活函数。

当基线模型优化结束后，其他模型将在此基础上进行部分微调，保留语音处理层和 encoder 层的模型参数，decoder 等模型参数则重新训练。所有实验在训练时共迭代 90 次，并且采用 warm-up 的学习率计划[9]，学习率 λ 的计算式如下：

$$\lambda = D^{-0.5} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5}) \tag{9}$$

其中，D 为模型维度，所有实验设置在优化的前 16000 步进行 warm-up。模型训练时采用 0.1 的标签平滑，并使用 Adam 优化器。训练结束后对后 20 代的模型参数进行平均获得最终模型。训练完成后的模型在预测阶段采用单步非自回归式的解码策略，因此所有实验均采用贪婪搜索解码。

4.3 实验结果

实验结果将对速度和准确度两方面的性能进行评估，准确度指标为字符错误率 CER。为保证公平性，对速度指标实时性因子 RTF 进行评估时，所有测试都在同一块 NVIDIA GeForce RTX 3080 Ti 显卡上进行，RTF 的计算式如下：

$$RTF = \frac{\text{模型总处理时长}}{\text{语音总时长}} \tag{10}$$

在表 2 中，我们首先将基线模型与提出的模型的实验结果进行比较。可以看出，提出的非自回归模型在加入音字特征转换模型后，基于 Bi-LSTM 的音字特征转换网络或深度 Bi-Transformer 音字特征转换网络的模型都实现了 CER 的下降，证明了本文提出的音字特征转换的有效性。其中，基于 Bi-Transformer 的特征转换模型效果提升最大，CER 从 9.0%减少到 8.3%，准确性相对提升了接近 7.8%，在本次实验中达到了最佳的性能。

表 2 实验结果比较

Table 2 Comparison of experimental results

模型		训练集 *CER*/%	实时性 *RTF*
基线模型			
CTC 帧信息合并		9.0	0.0020
提出的非自回归 Transformer			
+LSTM	1layer	8.9	0.0023
	2layer	9.0	0.0025
+Bi-LSTM	1layer	8.7	0.0046
	2layer	8.9	0.0050
+Bi-Transformer	残差连接	8.9	0.0021
	无残差连接	8.3	0.0028

对基于 LSTM 与基于 Bi-LSTM 的音字特征模型进行的实验可以发现，当两者采用 1 层结构时，其准确率都大于采用 2 层的结构，原因在于较深的 2 层结构在输入序列中提取出了更深的语义信息，即生成的隐向量包含了较多的语义特征，使得输出特征与所需求的字符特征存在差异，这对最终的输出字符特征产生了干扰，特征信息的错误提取最终导致了较低的识别准确率，该实验结果也与 3.4.1 小节中的分析相吻合。并且，采用 Bi-LSTM 结构的模型准确率高于 LSTM，原因在于 Bi-LSTM 双向连接的模式相比 LSTM 的单向连接模式可以更好地提取字符的上下文信息，输出向量中包含的语境信息使其能转换为更准确的字符特征，从而带来了更高的识别精度。尽管实现了错误率的降低，但是 Bi-LSTM 的解码速度与基线模型相比降低了 50%以上，实时性能较差，原因是 Bi-LSTM 中的基本神经单元是一种循环神经网络结构，串行计算过程产生了较大的延迟，这也导致了训练耗时更长、训练难度更大。

本文还对 Bi-Transformer 模型的连接方式进行了实验。在一些预训练语言模型中[24]，残差连接的方式有利于模型的优化与收敛，因此通过对模块之间的连接方式进行实验与对比，从而选择效果较好的连接方式。表 2 中的结果表明，无残差连接的模型具有更高的准确性，而采用残差连接时又导致了较高的错误率，分析该现象的原因是注意力网络具有较强的特征提取能力，残差连接的方式导致了特征信息的冗余与相互干扰，从而产生了较低的识别准确度。表 3 中列出了部分模型有(w/)、无(w/o)音字特征转换网络的识别结果。可以看出，当无音字特征转换网络的识别结果发生替换错误时，错误字符的读音与正确读音相近甚至相同，而进行特征转换

后，模型在读音特征中融合了上下文信息，输出特征因此包含了更多字符特征，从而获得了正确的转录结果，进一步验证了本文提出的音字特征转换的有效性。

表 3 部分识别结果

Table 3 Partial recognition results

基准	[zui4] 成为苹果近三个月以来股价下跌最严重的一次
w/o	[dui4] 成为苹果近三个月以来股价下跌对严重的一次
w/	[zui4] 成为苹果近三个月以来股价下跌最严重的一次
基准	[la1] 这个湖虽然没有漂浮的垃圾
w/o	[la1] 这个湖虽然没有漂浮的拉圾
w/	[la1] 这个湖虽然没有漂浮的垃圾
基准	[can1, sai4] 在参赛的二四支队伍中
w/o	[ta1, zai4] 在他在的二四支队伍中
w/	[can1, sai4] 在参赛的二四支队伍中

为进一步验证所提出的音字特征转换模型的有效性，以 Transformer 模型作为基线模型进行消融实验，结果如表 4 所列。当 Transformer 模型采用非自回归式解码后，解码方式的改变大幅提高了识别速度，而 CER 却有所降低。在此基础上，CTC 帧信息合并操作提升了部分性能，更进一步地，Bi-Transformer 音字特征转换模型的加入再一次提高了识别准确度，并且与基线模型相比在精度与速度方面均实现了较优的结果，进一步验证了音字特征转换模型的先进性。

表 4 消融实验对比

Table 4 Comparison of ablation experiments

模型	训练集 CER/%	实时性 RTF
Transformer	8.64	0.0312
+非自回归解码	10.96	0.0015
+CTC 帧信息合并	9.00	0.0020
+Bi-Transformer 音字特征转换	8.30	0.0028

我们还将本文提出的模型与 Transformer 模型和其他主流语音识别模型进行了对比，其中 Transformer 模型基于论文进行实验，还将其与以往工作中提出的采用自适应独立性假设的非自回归 AIANA-T 模型进行了对比，结果如表 5 所列。

表 5 与其他模型的结果比较

Table 5 Comparison with other models

模型	训练集 *CER*/%	实时性 *RTF*
Transformer	8.64	0.0312
LAS[28]	10.26	—
GRF[29]	9.48	—
Sync-Transformer[30]	8.91	0.1183
AIANA-T	8.79	0.0050
本文	8.30	0.0028

对比实验均在 AISHELL-1 数据集上进行，除 LAS[28] 使用基于 RNN 的特征提取网络外，其他模型均采用了基于自注意力机制的 encoder-decoder 架构。从表 5 可以看出，在无任何语言模型的情况下，本文提出的算法表现出了较强的竞争力，与其他模型相比，其 8.3% 的 CER 最小，与准确率相近的 Transformer 8.64% 的 CER 相比也下降了 4%，而 0.0028 的 RTF 与其他模型相比最小，是自回归模型 Transformer 的 11 倍，具有较快的解码速度。分析其原因是音字特征转换的加入在输出向量中融合了更多的字符特征，同时自注意力机制的架构与非自回归式的解码加快了解码的速度。以上结果表明，本文模型在速度与精度两方面均展现出了良好的语音识别性能。

结束语 本文提出了一个高性能的基于音字特征转换的非自回归 Transformer 端到端中文语音识别模型。在自回归语音识别 Transformer 的基础上，首先引入 CTC 帧信息合并实现了模型快速的非自回归式解码；其次利用提出的音字特征转换模型，在输出读音表征中融合更多的上下文信息，使输出表征包含更多字符特征，从而提高了转录文本的准确性。在中文数据集 AISHELL-1 上的实验结果表明，基于注意力机制的深度 Bi-Transformer 结构的表征模型具有很强的特征抽取能力，实现了字错误率 CER 8.3% 的识别准确度，并且其并行化的计算方式比 Bi-LSTM 具有更快的模型训练和解码速度，这些结果都验证了本文提出的非自回归 Transformer 模型快速而准确的语音识别效果，达到了可媲美当今主流语音识别模型的性能。未来我们将通过对语音表示计算和词表征模型的更深入的研究，进一步提升语音识别模型的性能。

参 考 文 献

[1] WANG H K, PAN J, LIU C. Research development and forecast of automatic speech recognition technologies[J]. Telecommunications Science, 2018, 34(2): 1-11.

[2] 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.

[3] ZHENG C J, WANG C L, JIA N. Survey of Acoustic Feature Extraction in Speech Tasks[J]. Computer Science, 2020, 47(5): 110-119.

[4] LI S, CAO F. Analysis and Trend Research of End-to-End Framework Model of Intelligent Speech Technology[J]. Computer Science, 2022, 49(6A): 331-336.

[5] AMODEI D, ANANTHANARAYANAN S, ANUBHAI R, et al. Deep speech 2: End-to-end speech recognition in english and mandarin[C]// International Conference on Machine Learning. PMLR. 2016: 173-182.

[6] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based large vocabulary speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 4945-4949.

[7] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 4960-4964.

[8] CHOROWSKI J K, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[J]. Advances in

Neural Information Processing Systems,2015,28:577-585.
[9] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. Advances in Neural Information Processing Systems,2017,30:5998-6008.
[10] GUO J,TAN X,HE D,et al. Non-autoregressive neural machine translation with enhanced decoder input[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33(1):3723-3730.
[11] DONG L,XU S,XU B. Speech-transformer:a no-recurrence sequence-to-sequence model for speech recognition[C]// 2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2018:5884-5888.
[12] SAK H,SENIOR A,BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[J]. arXiv:1402.1128,2014.
[13] CHUNG J,GULCEHRE C,CHO K H,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv:1412.3555,2014.
[14] GRAVES A. Generating sequences with recurrent neural networks[J]. arXiv:1308.0850,2013.
[15] CHEN X,ZHANG S,SONG D,et al. Transformer with bidirectional decoder for speech recognition[J]. arXiv:2008.04481,2020.
[16] BAI Y,YI J,TAO J,et al. Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from bert[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2021,29:1897-1911.
[17] HIGUCHI Y,INAGUMA H,WATANABE S,et al. Improved mask-CTC for non-autoregressive end-to-end ASR[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2021:8363-8367.
[18] LI J,WANG X,LI Y. The speech transformer for large-scale mandarin chinese speech recognition[C]// 2019 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP 2019). IEEE,2019:7095-7099.
[19] ZHOU P,FAN R,CHEN W,et al. Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding[J]. arXiv:1911.00203,2019.
[20] DO M N,LU Y M. Multidimensional filter banks and multiscale geometric representations[J]. Foundations and Trends in Signal Processing,2012,5(3):157-264.
[21] 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. 2006:369-376.
[22] MIAO H,CHENG G,GAO C,et al. Transformer-based online CTC/attention end-to-end speech recognition architecture[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2020:6084-6088.
[23] PETERS M E,NEUMANN M,IYYER M,et al. Deep Contextualized Word Representations[J]. arXiv:1802.05365,2018.
[24] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[J]. arXiv:1810.04805,2018.
[25] BU H,DU J,NA X,et al. Aishell-1:An open-source mandarin speech corpus and a speech recognition baseline[C]// 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment(O-COCOSDA). IEEE,2017:1-5.
[26] OTT M,EDUNOV S,GRANGIER D,et al. Scaling Neural Machine Translation[J]. arXiv:1806.00187,2018.
[27] PARK D S,CHAN W,ZHANG Y,et al. Specaugment:A simple data augmentation method for automatic speech recognition[J]. arXiv:1904.08779,2019.
[28] SHAN C,WENG C,WANG G,et al. Component fusion:Learning replaceable language model component for end-to-end speech recognition system[C]// ICASSP 2019-2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2019:5361-5635.
[29] FAN C,YI J,TAO J,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.
[30] TIAN Z,YI J,BAI Y,et al. Synchronous transformers for end-to-end speech recognition[C]// 2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP 2020). IEEE,2020:7884-7888.

TENG Sihang, born in 1996, postgraduate. His main research interests include deep learning and speech recognition.

WANG Lie, born in 1969, professor, master supervisor. His main research interests include deep learning, image processing and FPGA.

(责任编辑:何杨)