计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009718] [CSTR: 32024.14.csa.009718]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Conformer-SE 的端到端语音识别①

马永杰[1], 李　罡[2]

[1](吉林化工学院 信息与控制工程学院, 吉林 132022)
[2](白城师范学院 机械与控制工程学院, 白城 137000)
通信作者: 李　罡, E-mail: ligang@bcnu.edu.cn

摘　要: 基于自注意力机制的 Transformer 端到端模型在语音识别任务中表现出了卓越的性能. 然而, 该模型在浅层处理时对局部特征信息的捕捉能力存在一定的局限, 同时也没有充分考虑不同块之间的相互依赖性. 为了解决这些问题, 提出了一种改进的 Conformer-SE 端到端语音识别系统模型. 该模型首先采用了 Conformer 结构来替代 Transformer 中的编码器部分, 从而增强了模型对局部特征的提取能力. 接着, 通过引入 SE 注意力通道机制, 将每个块的输出以加权求和的形式整合到最终的输出中. 在 Aishell-1 这一公开数据集上的实验结果显示, 相较于原始的 Transformer 模型, Conformer-SE 模型在字符错误率上相对降低了 18.18%.

关键词: 语音识别; 端到端; Transformer; Conformer; SE 注意力通道

引用格式: 马永杰,李罡.基于 Conformer-SE 的端到端语音识别.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9718.html

End-to-end Speech Recognition Based on Conformer-SE

MA Yong-Jie[1], LI Gang[2]

[1](School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin 132022, China)
[2](School of Mechanical and Control Engineering, Baicheng Normal University, Baicheng 137000, China)

Abstract: The end-to-end Transformer model based on the self-attention mechanism shows superior performance in speech recognition. However, this model has limitations in capturing local feature information during shallow processing and does not fully consider the interdependence between different blocks. To address these issues, this study proposes Conformer-SE, an improved end-to-end model for speech recognition. The model first adopts the Conformer structure to replace the encoder in the Transformer model, thus enhancing its ability to extract local features. Next, by introducing the SE channel attention mechanism, it integrates the output of each block into the final output through a weighted sum. The experimental results on the Aishell-1 dataset show that the Conformer-SE model reduces the character error rate by 18.18% compared to the original Transformer model.

Key words: speech recognition; end-to-end; Transformer; Conformer; SE attention channel

随着人工智能技术的快速发展, 语音识别技术[1]早已转换成产品, 并被广泛应用于客服电话、医疗保健、出行驾驶、智能家居等各种场景, 已经成为人们日常生活和工作中不可或缺的一部分. 此外, 随着边缘计算和物联网技术的融合, 语音识别技术正逐渐向更高效、更个性化的方向发展, 为用户提供更加无缝和智能的交互体验.

早在 20 世纪 50 年代, 贝尔实验室就开发了名为 Audrey 的孤立词语音识别系统[2]. 到 60 年代末, 苏联科学家 Vintsyuk 提出了使用动态规划算法 (dynamic time warping, DTW) 来解决语音识别中的时间对齐问题[3], 这一技术进一步加快了语音识别的发展. 从 20 世

① 基金项目: 2022 年度吉林省教育厅科学技术研究项目 (JJKH20220013KJ); 2023 年大学生创新创业训练计划 (202310206035)
收稿时间: 2024-05-28; 修改时间: 2024-06-26; 采用时间: 2024-07-11; csa 在线出版时间: 2024-10-31

纪 80 年代开始, 隐马尔可夫模型 (hidden Markov model, HMM) 在语音识别领域中得到运用[4], 语音识别技术逐渐从孤立词的识别阶段迈向连续词的识别阶段. 后来, 通过与高斯混合模型 (Gaussian mixture model, GMM) 相结合[5], 基于 HMM 的语音识别模型的性能得到了提升.

21 世纪初, 深度学习的概念被提出, 人工神经网络 (ANN) 被首次用于语音识别[6], 但由于当时的实验环境与理论不足等原因, 基于人工神经网络的语音识别并没有发展起来. 直到 2011 年, Dahl 等人[7]提出用音素作为建模单位, 将深度神经网络 DNN (deep neural network) 与 HMM 相结合, 与 GMM-HMM 相比, 深度神经网络-隐马尔可夫模型 (DNN-HMM) 的引入显著提升了语音识别的性能, 使其达到了真实用户可接受的水平[8].

近年来, 为了简化语音识别系统的结构和提高效率, 端到端语音识别技术应运而生[9]. 端到端语音识别是一种直接从语音信号到文本输出的完整语音识别系统, 它将传统语音识别系统中的多个组件整合为一个统一的模型[10]. 目前端到端语音识别方法[11]主要分 3 类: (1) 连接时序分类 (connectionist temporal classification, CTC) 方法[12]; (2) 循环神经网络变换器 (recurrent neural network transducer, RNN-T) 方法[13]; (3) 基于注意力机制 (attention) 的方法[14]. 近年来, 随着基于自注意力机制的 Transformer[15]模型在文本翻译领域取得了竞争性的结果后, Dong 等人[16]首次将 Transformer 模型应用于自动语音识别领域中, 并取得了比较好的效果. 然而, 相对于自注意力机制, Transformer 模型在提取局部特征方面的能力稍显不足, 而卷积网络则更擅长于提取局部特征. 为了加强提取特征的能力, Gulati 等人[17]提出了 Conformer 模型, Conformer 模型通过将卷积模块集成到 Transformer 编码器中, 有效地提升了局部特征提取的能力. 为了进一步研究局部和全局特征的关系, Peng 等人[18]提出了一种 Branchformer 模型. 但是, 这些模型都没有考虑到每个块之间的相互依赖关系, 为了提高每个块之间的相互依赖关系, 本文提出了一种改进的 Conformer-SE 语音识别模型. 在语音识别任务中, 我们用 Conformer 结构替代了 Transformer 的编码器部分, 并通过残差连接将编码器和解码器中的每个块的输出结合 SE 注意力通道机制[19]进行加权求和, 以得到最终的输出. 在公开数据集 Aishell-1 上, 我们通过与基线模型进行对比实验, 评估了语音识别的字符错误率和实时率. 结果显示, 在实时率几乎未受影响的情况下, 我们的模型将字符错误率降低了 18.18%, 这一结果证明了该模型的先进性. 此外, 我们提出的 SE 加权求和结构在基线模型上的应用也得到了验证, 结果表明该结构能够提升基线模型的准确率, 进一步展示了该模型具有良好的泛化能力.

1 基线模型

Speech Transformer[20]模型是本文采用的基线模型, 该模型基于 Transformer 架构, 并在自动语音识别 (automatic speech recognition, ASR) 和语音合成 (text to speech, TTS) 等领域表现出色. Speech Transformer 的核心由 3 个主要部分构成: 编码器 (encoder)、解码器 (decoder) 以及位置编码 (positional encoding). 编码器和解码器是模型的核心[21], 它们由多个重复的层叠加而成, 每层都融合了多头自注意力 (multi-head self-attention) 和前馈神经网络 (feed-forward network). 多头自注意力机制的引入使模型可以更好地捕捉输入序列中的长距离依赖关系[22], 而前馈神经网络则通过非线性变换增强了模型的表征能力. 鉴于 Transformer 架构本身不具有处理序列顺序的能力, 位置编码的引入至关重要. 位置编码为序列中的每个元素提供位置信息, 帮助模型理解和处理时序数据. 在 Speech Transformer 中, 位置编码被叠加到编码器和解码器的输入特征上, 从而使模型能够捕捉并利用这些位置信息, 其模型结构图如图 1 所示.

尽管 Speech Transformer 在语音识别领域已经显示出其强大的性能, 但它仍然面临着一些挑战. 首先, Speech Transformer 主要依赖于全局自注意力机制, 而这种机制可能忽视了语音信号中至关重要的局部特征信息. 其次, Transformer 模型中的各层直接堆叠可能导致在训练过程中部分关键信息的丢失, 同时各个块之间的相互依赖关系不足. 为了克服这些问题, 本文在第 2 节中提出了相应的改进策略.

1.1 多头注意力

注意力机制是一种重要的深度学习技术[23], 它使得模型能够选择性地关注于输入序列中的关键部分, 并据此分配不同的权重. 在 Transformer 模型中, 自注意力机制是其核心构件, 该机制通过缩放点积注意力来实现. 缩放点积注意力的模型结构如图 2 所示.

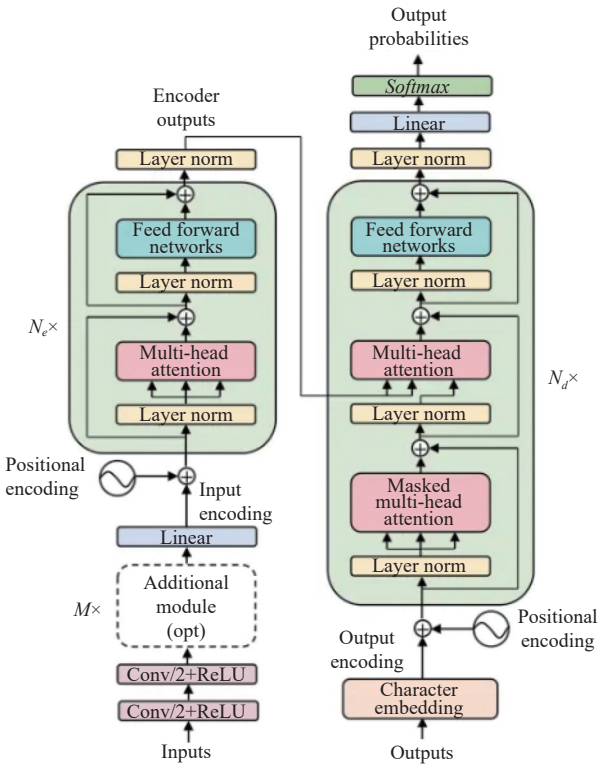

图 1 Speech Transformer 结构图

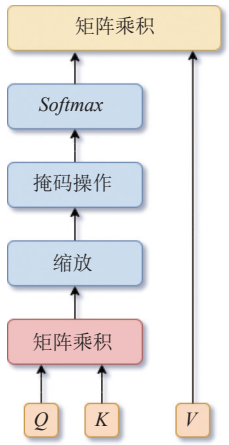

图 2 缩放点注意力机制

缩放点注意力机制有效提升了计算的效率和模型的性能. 首先, 通过计算查询向量 Q 与键向量 K 的点积, 获得原始的注意力得分. 接着, 为了防止数值过大而导致的稳定性问题, 这些得分会被除以一个与查询向量维度相关的缩放因子进行调整. 随后, 经过 *Softmax* 函数处理[24], 这些缩放后的得分转化为注意力权重. 最后, 将这些权重应用于值向量 V, 通过权重与值向量的乘积, 得到了加权值向量. 整个过程可以用式 (1) 描述:

$$attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, $Q\in R^{t_q\times d_q}$, $K\in R^{t_k\times d_k}$, $V\in R^{t_v\times d_v}$; t 表示输入数量; d 表示特征维度大小.

多头注意力机制是由多个缩放点积注意力的基础单元堆叠而成 (如图 3 所示), 其计算过程如下: 首先将查询 Q、键 K 和值 V 分别映射到 d_q, d_k 和 d_v 的维度, 切分为 h 个头; 再将各个子空间的输出进行 *concat* 拼接, 最后经过一个可学习的线性映射得到输出, 如式 (2) 和式 (3) 所示:

$$head_i=attention(QW_i^Q,KW_i^K,VW_i^V) \tag{2}$$

$$MultiHead(Q,K,V)=concat(head_1,\cdots,head_h)W^O \tag{3}$$

其中, $W_i^Q\in R^{d_m\times d_q}$, $W_i^K\in R^{d_m\times d_k}$, $W_i^V\in R^{d_m\times dv}$ 和 $W^O\in R^{d_m\times d_m}$ 是可训练的参数矩阵; h 是注意力头的数量; d_m 是输出的特征维度大小, 本文设定 $d_q=d_k=d_v=d_m/h$.

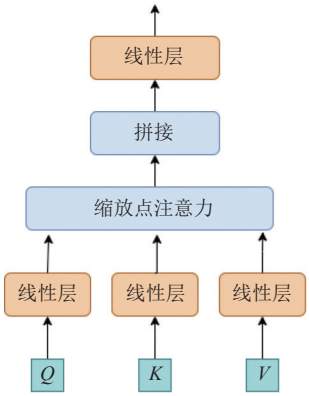

图 3 多头注意力机制

1.2 位置编码

传统的循环神经网络 (RNN) 和长短时记忆网络 (LSTM) 等模型能够通过其循环结构自然地捕捉序列中元素的位置信息[25]. 然而, Transformer 模型是一种基于自注意力机制的模型, 它不具有循环结构, 因此无法直接从输入序列中推断出元素的位置顺序. 为了解决这一问题, 位置编码被引入到 Transformer 模型中, 即确定每个单词之间的距离和位置, 其计算方法如下:

$$PE_{(pos,2i)}=\sin\left(\frac{pos}{10000^{\frac{2i}{d_m}}}\right) \tag{4}$$

$$PE_{(pos,2i+1)}=\cos\left(\frac{pos}{10000^{\frac{2i}{d_m}}}\right) \tag{5}$$

其中, *pos* 表示当前单词的位置, *PE* 表示位置 *pos* 上的

编码值, $2i$ 或者 $2i$+1 代表位置编码向量的一个分量, $2i$ 代表偶数, $2i$+1 代表奇数, d_m 表示是输出的特征维度大小.

1.3 前馈神经网络

在 Transformer 结构中, 前馈神经网络 (feed-forward neural network, *FFN*) 是一个重要的组件, 它位于 Transformer 的每个编码器和解码器层之后. 前馈神经网络的主要作用是对嵌入向量进行非线性变换和映射, 以强化位置信息的表示和特征的提取. 前馈神经网络架构包含两个线性层 (即全连接层), 其间嵌入了一个非线性激活函数. 在 Transformer 中, 前馈神经网络可以表示为式 (6):

$$FFN(X)=\max(0,XW_1+b_1)W_2+b_2 \tag{6}$$

其中, X 是输入张量, $W_1\in R^{d_m\times d_{\text{middle}}}$, $W_2\in R^{d_m\times d_{\text{middle}}}$, $b_1\in R^{d_{\text{middle}}}$ 和 $b_2\in R^{d_m}$ 是可学习的权重矩阵和偏置向量, d_{middle} 是中间层的特征维度大小.

2 Conformer-SE 模型

为了更有效地提取局部特征信息并提升模型的识别准确性, 本研究将含有卷积模块的 Conformer 结构替换 Transformer 中的编码器部分. 在 Conformer 模型中, 我们采用了相对位置编码策略, 与传统的绝对位置编码相比, 它能更灵活地适应序列长度的变化. 此外, 针对基线模型直接堆叠而导致损失重要局部信息, 缺乏各个块之间的相互依赖关系, 本研究还提出了一种新颖的压缩激励模型 (squeeze-and-excitation, SE) 方法, 该方法通过建立编码器与解码器中各输出模块之间的联系, 强化了模块间的相互作用. 最终, 通过加权求和机制, 实现了模块集成输出的优化. 这种结构不仅增强了模型对序列中关键信息的捕捉能力, 也为处理变长序列提供了更加稳健的解决方案. 其整体模型结构如图 4 所示. 相比于基线模型中的直接堆叠结构, 该模型首先通过残差网络将每一个 block 的输出都分别提取出来压缩存储在 SE 通道符中, 再将各个块的输出通过相加得到块集成的最终结果.

2.1 Conformer 编码器

Conformer 的核心思想是在 Transformer 的基础上加入卷积层, 以此来更好地捕捉语音信号中的局部特征和长距离依赖关系. 具体来说, Conformer 结构包括一系列的处理步骤, 如 SpecAug 数据增强[26]、Convolution subsampling 用于特征提取和降维、多个 Conformer block 用于处理特征, 以及最终的分类器. Conformer block 内部包含了自注意力模块、卷积模块和前馈网络, 这些模块的组合使得 Conformer 能够同时捕捉到局部和全局的信息, 从而在语音识别任务中取得了显著的性能提升. 图 5 左侧展示了 Conformer 模型的整体结构, 右侧展示了 Conformer block 具体组成.

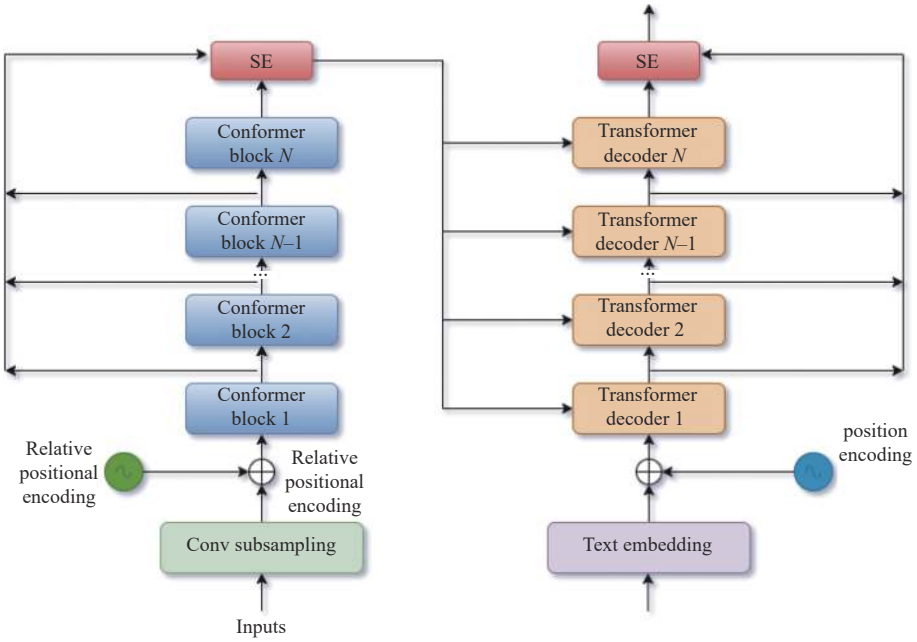

图 4 Conformer-SE 模型结构

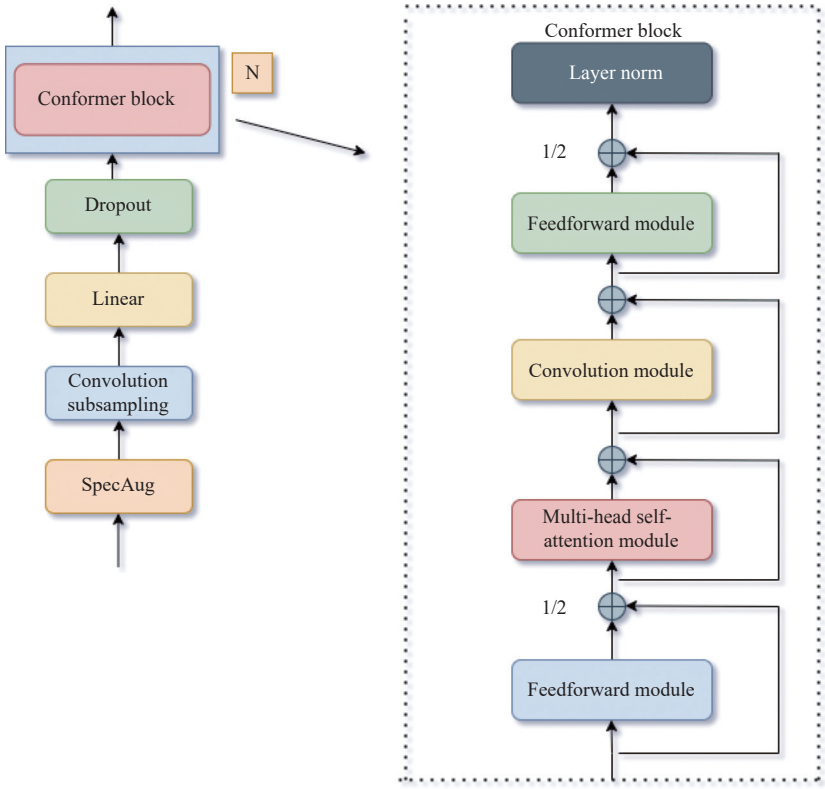

图 5 Conformer 模型结构

在 Conformer block[27]中, 它在首末两端均设置了一个输出为一半的前馈神经网络, 中间夹着一个多头注意力模块和一个卷积模块. 具体到数学表达上, 对于第 i 个块接收到的输入 x_i, 输出为 h_i 的计算公式如下:

$$\begin{cases} \tilde{x}_i = x_i + \dfrac{1}{2}FFN(x_i) \\ x'_i = \tilde{x} + MHSA(\tilde{x}_i) \\ x''_i = x'_i + Conv(x'_i) \\ h_i = Layernorm\left(x''_i + \dfrac{1}{2}FFN(x''_i)\right) \end{cases} \tag{7}$$

其中, *FFN* 表示前馈神经网络模块, *MHSA* 表示多头注意力模块, *Conv* 表示卷积模块, *Layernorm* 代表归一化模块.

2.2 相对位置编码

相对位置编码是自注意力模型中的一种关键位置编码技术, 它通过精确捕捉序列元素间的相对位置关系, 极大地增强了模型对序列内不同位置关联性的理解能力. 这种方法是 Transformer-XL 架构[28]的核心组成部分, 它相较于传统的绝对位置编码, 提供了更高的灵活性和适应性. 通过对相对距离的有效编码, 相对位置编码能够优化模型对序列任务的处理效果, 使其在各种应用场景中展现出更卓越的性能.

2.3 SE 通道注意力机制

SE 通道注意力机制是一种先进的神经网络架构改进方法, 旨在通过显式地建模通道间的依赖关系来增强网络的表征能力. 这种机制最初由 Hu 等人在文献[29]中提出, 并迅速成为提升深度学习模型性能的重要手段之一. SE 模块的核心思想是对特征通道进行动态的重校准, 以便突出重要的特征并抑制不太重要的特征.

SE 通道注意力机制主要包含两个步骤: Squeeze (压缩) 和 Excitation (激励).

Squeeze 步骤: 通过执行全局平均池化, 将每个特征通道的信息汇总为一个单一的数值. 这个过程有助于模型捕获全局信息, 为接下来的重校准步骤提供基础.

Excitation 步骤: 以 Squeeze 步骤的输出作为输入, 通过一个由两个全连接层组成的小型网络 (其中包含一个 ReLU 激活[30]层和一个 Sigmoid 激活层), 学习到每个通道的重要性权重. 这一步的输出是一个在 0–1 之间的权重系数, 用于调整每个通道的特征响应.

最后, 原始特征图通过与这些学习到的权重系数

相乘, 实现了特征的动态重校准. 这种机制使得模型能够侧重于更加重要的特征, 从而提升了模型对信息的利用效率和最终的性能.

SE 通道注意力模型结构如图 6 所示.

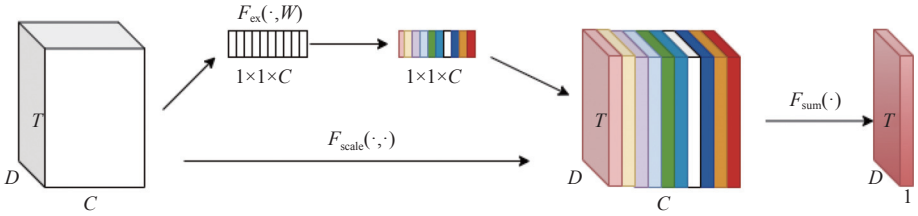

图 6 SE 通道注意力机制结构

Conformer 架构虽然强大, 但并未充分强调不同块之间的相互关联. 为了弥补这一不足, 我们采取了一种方法: 从每个块的输出特征中提取关键信息, 并将其整合到单一通道中, 从而捕捉到全局的输出信息. 接着, 我们运用全局平均池化操作来压缩这些信息, 形成一个通道数据. 最终, 我们通过加权求和的方式, 结合这些通道数据, 生成了更为综合的块集成输出, 从而增强了模型处理序列任务时的整体表现. 其计算过程如下:

$$z_c = F_{sq}(y_c) = \frac{1}{T \times D}\sum_{i=1}^{T}\sum_{j=1}^{D} y_c(i,j) \tag{8}$$

其中, $y_c \in R_{T\times D}$指的是第 c 个块的输出, T 和 D 是维度.

$$s = F_{ex}(z,W) = \sigma(W_2\delta(W_1 z)) \tag{9}$$

其中, σ为 Sigmoid 函数, δ 为 ReLU 函数, $W_1, W_2 \in R^{c\times\frac{c}{r}}$, c 为 channel 的通道数, r 为缩放因子. 这里我们设置的 $c_{encoder}$=$c_{decoder}$=6, r=1.

$$\tilde{y}_c = F_{scale}(y_c, s_c) = s_c y_c \tag{10}$$

$$\tilde{y} = F_{sum}(\tilde{y}_c) = \sum_{c=1}^{N}\tilde{y}_c \tag{11}$$

其中, $\tilde{y}_c$指的是第 c 个通道的输出, $\tilde{y}$代表的是最终的块集成输出.

2.4 Transformer 解码器

解码器是沿用了基线模型中图 1 右侧部分, 相较于左侧的编码器部分, 解码器模块中添加了一个交叉注意模块 (*MHCA*). 具体到数学表达上, 对于第 i 个块接收到的输入 x_i, 输出为 y_i 的计算公式如下:

$$x_i' = x_i + MHSA(x_i) \tag{12}$$

$$x_i'' = x_i' + MHCA(x_i', \tilde{y}) \tag{13}$$

$$y_i = Layernorm\left(x_i'' + FFN(x_i'')\right) \tag{14}$$

其中, $\tilde{y}$表示编码器块集成的输出.

3 实验及结果

3.1 实验数据和实验环境

本研究采用了中文普通话数据集 Aishell-1 进行实验, 这是一套由北京科技创新研究院 (Beijing Institute of Science and Technology Innovation) 公开发布的中文语音识别数据集. Aishell-1 数据集是为自动语音识别 (ASR) 技术的研究与开发量身打造的, 总计包含约 178 h 的高质量中文普通话语音数据, 覆盖了超过 400 名不同说话人, 其中男女发音者的比例接近均衡. 该数据集囊括了来自各个领域的语句, 包括日常交流、财经新闻、体育报道等, 为语音识别的训练与测试提供了一个多元化的平台. Aishell-1 的录音在静谧环境下完成, 以确保高品质的音频效果, 所有的语音文件均以 16 kHz 的采样率和 16 位的深度, 以单声道 WAV 格式进行存储. 这个数据集不仅向研究界开放, 同时也支持商业应用, 对推动中文语音识别技术的进步贡献了重要资源.

在本研究中, 实验环境搭建在搭载 Python 3.8 的 Ubuntu 18.04 操作系统上. 计算核心采用了 Intel(R) Xeon(R) Silver 4214R CPU, 该处理器主频为 2.4 GHz, 搭配了 90 GB 的系统内存. 语音处理方面, 实验使用了一块具备 12 GB 显存的 RTX 3080 Ti GPU. 此外, 实验采用了 PyTorch 1.7.0 作为深度学习的主要框架, 并结合了 Cuda 11.0 技术以优化 GPU 的性能表现.

3.2 评价指标

本文在 Aishell-1 数据集上的实验分析, 采用字符错误率 (character error rate, *CER*) 作为评价标准, 其计算公式如下:

$$CER = \frac{D+S+I}{N} \tag{15}$$

其中, D 表示删除错误的字数量, S 表示替换错误的字数量, I 表示插入错误的数量, N 是参考文本中字符的总数.

对于模型的推理速度评价指标采用的是实时率(real time factor, *RTF*), 其计算公式如下:

$$RTF = \frac{T_{\text{ASR}}}{T} \tag{16}$$

其中, T_{ASR} 表示解码时长, T 是指音频时长, 实时率越小, 表示推理速度和解码速度越快.

3.3 训练过程

首先, 对数据进行预处理. 在本文中, 音频输入特征采用帧长为 25 ms, 帧移为 10 ms, 特征维度为 80 的 Fbank 特征[31]. 此外, 为提升模型的鲁棒性, 对语音信号的频谱特征应用了 SpecAugment 技术, 并在编码器的输入前进行了卷积下采样处理. 编码器和解码器均采用了 N 为 6 的块结构进行实验, 每个模块的输出维度设为 256, 注意力头数为 8, 前馈神经网络隐藏维度为 2 048.

接着, 将每个块结构的输出通过 SE 通道注意力机制进行加权求和, SE 的通道数设为 6, 以存储 6 个块结构的输出.

最后, 在训练阶段, 设置 batch_time 为 150, 采用累计梯度 accum_grad 为 8, 训练轮次 epoch 为 80. 采用动态学习率策略, 将热身步数 warmup_steps 设为 16 000, 最高学习率 lr 为 0.000 4. 优化器选择 Adam. 训练结束后, 对最后 10 个模型进行参数平均得到最终的训练模型.

3.4 实验结果

在数据集 Aishell-1 上进行的实验结果, 如表 1 所示, 在未引入 SE 通道注意力机制的情况下, 将基于 Speech Transformer 的语音识别系统 (基线模型) 与 Conformer 模型进行了对比. 观察到, 尽管两者在相同数据集上接受训练, 但 Conformer 模型展现出了更优越的性能, 在测试集上其字错误率显著低于 Speech Transformer 模型, 实现了 12.12% 的相对降低. 这表明 Conformer 模型在语音识别任务中具有更强的鲁棒性和准确性. 同时, 我们还将 SE 通道注意力机制块集成只用到 Transformer 和 Conformer 的编码器上进行了测试, 观察到 Transformer 和 Conformer 在测试集上的字符错误率分别下降了 0.2% 和 0.1%.

接着, 在 Aishell-1 数据集中, 通过引入 SE 通道注意力机制, 将编码器与解码器同时块集成, 我们实施了图 4 所提出的方法. 根据表 2 所示的结果, 可以明显观察到, Speech Transformer 模型和 Conformer 模型在测试集上的字符错误率分别相对减少了 7.57% 和 6.89%. 而 Conformer-SE 模型相比于基线模型在测试集上的字错误率更是相对降低了 18.18%. 这一显著的性能提升充分证明了所加 SE 模型的有效性及其在扩展性方面的表现.

表 1 Aishell-1 数据集上的 *CER* (%)

语音识别系统	验证集*CER*	测试集*CER*
Speech Transformer	6.00	6.60
Conformer	5.30	5.80
Speech Transformer-encoder SE	5.90	6.40
Conformer-encoder SE	5.20	5.70

表 2 添加 SE 在 Aishell-1 数据集上的 *CER* (%)

语音识别系统	验证集*CER*	测试集*CER*
Transformer-SE	5.70	6.10
Conformer-SE	5.10	5.40

本文还将 Conformer-SE 模型的性能与先前研究中基于 Transformer 模型[32,33]在 Aishell-1 数据集上的成果进行了比较, 具体见表 3. 结果显示, 相较于其他先前的工作, Conformer-SE 模型展现了更优异的性能表现.

表 3 Aishell-1 数据集上不同模型的 *CER* (%)

模型	验证集*CER*	测试集*CER*
Sync-Transformer	7.91	8.91
LAS	—	10.56
RNN-T	10.13	11.82
ESPNET	6.00	6.70
Conformer-SE	5.10	5.40

语音识别的实时率是衡量语音识别速度的一个重要指标, 该值越低, 意味着处理单位时间内语音数据的能力越强, 效率越高. 根据表 4 的测试结果可以看出, Conformer 模型相较于 Speech Transformer 模型, 实时率升高了 0.001 2, 而 Conformer-SE 模型也仅比 Speech Transformer 模型高了 0.001 6. 综合评估, 虽然 Conformer-SE 模型在实时率上略低于原模型, 但其识别率显著提高, 综合考虑两方面的性能, Conformer-SE 模型展现出了一定的先进性.

表 4 语音识别的实时率

语音识别系统	*RTF*
Speech Transformer	0.064 8
Conformer	0.066 0
Transformer-SE	0.065 0

Conformer-SE	0.0664

4 结束语

在这项研究中, 本文提出了一种基于 Conformer-SE 的端到端语音识别系统. Conformer 中的卷积模块能更好地捕捉细粒度的局部特征信息, 弥补了 Transformer 的不足. 同时, SE 通道注意力机制使输出能够考虑每个块的输出, 提高了每个块之间的相互依赖性. 通过在 Aishell-1 中文普通话数据集上的测试发现: 通过残差的方式将每个块的输出结果以 SE 通道加权求和, 可以更好地提高模型的语音识别能力. 此外, 本文提出的模型在 Transformer 上也表现不错, 具有很好的泛化能力. 与之前发表的其他基于 Transformer 模型相比, 具有一定的先进性. 然而, 在实验中我们仍沿用了 Transformer 中的解码器结构, 未来的研究将重点放在解码器上, 以进一步提高语音识别性能.

参考文献

1 王家, 龙冬梅. 深度学习在语音识别中的应用综述. 电脑知识与技术, 2020, 16(34): 191–192, 197.

2 Davis KH, Biddulph R, Balashek S. Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 1952, 24(6): 637–642. [doi: 10.1121/1.1906946]

3 Vintsyuk TK. Speech discrimination by dynamic programming. Cybernetics, 1968, 4(1): 52–57.

4 吴延占. 基于 HMM 与遗传神经网络的改进语音识别系统. 计算机系统应用, 2016, 25(1): 204–208.

5 Karpagavalli S, Chandra E. Phoneme and word based model for Tamil speech recognition using GMM-HMM. Proceedings of the 2015 International Conference on Advanced Computing and Communication Systems. Coimbatore: IEEE, 2015. 1–5.

6 张丹. 深度学习神经网络在语音识别中的应用探讨. 电子世界, 2021(6): 67–68.

7 Dahl GE, Yu D, Deng L, *et al*. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011. 4688–4691.

8 Dahl GE, Yu D, Deng L, *et al*. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30–42. [doi: 10.1109/TASL.2011.2134090]

9 李荪, 曹峰. 智能语音技术端到端框架模型分析和趋势研究. 计算机科学, 2022, 49(6A): 331–336.

10 Zhang QL, Chen JF, Bai JS. Language model based non-speech recognition method. Proceedings of the 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Dalian: IEEE, 2019. 1–5.

11 王澳回, 张珑, 宋文宇, 等. 端到端流式语音识别研究综述. 计算机工程与应用, 2023, 59(2): 22–33.

12 Miao HR, Cheng GF, Zhang PY, *et al*. Online hybrid CTC/attention architecture for end-to-end speech recognition. Proceedings of Interspeech. Graz: Interspeech, 2019. 2623–2627.

13 Zhang Q, Lu H, Sak H, *et al*. Transformer transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 7829–7833.

14 Chorowski J, Bahdanau D, Serdyuk D, *et al*. Attention-based models for speech recognition. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 28.

15 Chang XK, Zhang WY, Qian YM, *et al*. End-to-end multi-speaker speech recognition with Transformer. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6134–6138.

16 Dong LH, Xu B. CIF: Continuous integrate-and-fire for end-to-end speech recognition. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6079–6083.

17 Gulati A, Qin J, Chiu CC, *et al*. Conformer: Convolution-augmented Transformer for speech recognition. Proceedings of the Interspeech. Shanghai: Interspeech, 2020. 5036–5040.

18 Peng YF, Dalmia S, Lane I, *et al*. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022. 17627–17643.

19 徐沁, 梁玉莲, 王冬越, 等. 基于 SE-Res2Net 与多尺度空谱融合注意力机制的高光谱图像分类. 计算机辅助设计与图形学学报, 2021, 33(11): 1726–1734.

20 Zhao YZ, Ni CJ, Leung CC, *et al*. Universal speech Transformer. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: Curran Associates Inc., 2020. 5021–5025.

21 HeZR, Shen QF, Wu JX, et al. Transformer encoder-based

multilevel representations with fusion feature input for speech emotion recognition. Journal of Southeast University (English Edition), 2023, 39(1): 68–73.

22 王明. 基于元学习的多头注意力时序卷积的入侵检测. 网络安全与数据治理, 2023, 42(7): 49–54.

23 张英, 拥措, 于韬. 基于动态多头注意力机制的藏文语言模型. 计算机工程与设计, 2023, 44(12): 3707–3713.

24 钟雨昂, 袁伟伟, 关东海. 基于 Softmax 的加权 Double Q-Learning 算法. 计算机科学, 2024, 51(6A): 230600235. [doi: 10.11896/jsjkx.230600235]

25 翁鸣昊, 项兴华, 陈俊涛, 等. 基于 LSTM 与 Transformer 的大坝变形预测研究. 中国农村水利水电, 2024(4): 250–257.

26 崔晨露, 崔琳. 面向数据增强的轻量化语音情感识别. 计算机与现代化, 2023(4): 83–89, 100. [doi: 10.3969/j.issn.1006-2475.2023.04.013]

27 陈戈, 谢旭康, 孙俊, 等. 使用 Conformer 增强的混合 CTC/Attention 端到端中文语音识别. 计算机工程与应用, 2023, 59(4): 97–103. [doi: 10.3778/j.issn.1002-8331.2111-0462]

28 Dai ZH, Yang ZL, Yang YM, *et al*. Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 2978–2988.

29 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

30 刘霞, 王迪. 深度 ReLU 神经网络的万有一致性. 中国科学: 信息科学, 2024, 54(3): 638–652.

31 崔琳, 王芷悦. 基于 LFBank 与 FBank 混合特征的声纹识别研究. 计算机科学, 2022, 49(11A): 211000194. [doi: 10.11896/jsjkx.211000194]

32 Chan W, Jaitly N, Le Q, *et al*. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016. 4960–4964.

33 Tian ZK, Yi JY, Bai Y, *et al*. Synchronous Transformers for end-to-end speech recognition. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 7884–7888.

(校对责编: 张重毅)