计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):194−202 [doi: 10.15888/j.cnki.csa.008638]

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

结合 Conformer 与 N-gram 的中文语音识别①

许鸿奎[1,2], 卢江坤[1], 张子枫[1], 周俊杰[1], 胡文烨[1], 姜彤彤[1], 郭文涛[1], 李振业[1]

[1](山东建筑大学 信息与电气工程学院, 济南 250101)
[2](山东省智能建筑技术重点实验室, 济南 250101)
通信作者: 许鸿奎, E-mail: xhkui2009@163.com

摘　要: Transformer 模型对输入序列中重要的信息进行学习, 相比传统的 ASR (automatic speech recognition) 模型提升了准确性. Conformer 模型在 Transformer 的编码器中加入卷积模块, 增加了获取细微局部信息的能力, 进一步提高了模型性能. 本文结合使用 Conformer 模型和 N-gram 语言模型 (language model , LM) 用于中文语音识别, 获得了良好的识别效果. 在数据集 AISHELL-1 和 aidatatang_200zh 上的实验表明, 使用 Conformer 模型字错率分别可降低到 5.79% 和 5.60%, 较 Transformer 模型降低了 5.82% 和 2.71%. 结合 N-gram 语言模型后字错率分别可降低到 4.86% 和 5.10% 达到最佳性能, 实时率 (real time factor , RTF) 达到 0.145 66. 测试信噪比降低为 20 dB 时模型字错率才明显下降到 8.58%, 表明该模型具有一定的抗噪能力.

关键词: 语音识别; Transformer; 语言模型; Conformer; 深度学习

引用格式: 许鸿奎,卢江坤,张子枫,周俊杰,胡文烨,姜彤彤,郭文涛,李振业.结合 Conformer 与 N-gram 的中文语音识别.计算机系统应用,2022,31(7):194–202. http://www.c-s-a.org.cn/1003-3254/8638.html

Chinese Speech Recognition Based on Conformer and N-gram

XU Hong-Kui[1,2], LU Jiang-Kun[1], ZHANG Zi-Feng[1], ZHOU Jun-Jie[1], HU Wen-Ye[1], JIANG Tong-Tong[1], GUO Wen-Tao[1], LI Zhen-Ye[1]

[1](School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China)
[2](Shandong Provincial Key Laboratory of Intelligent Building Technology, Jinan 250101, China)

Abstract: The Transformer model can learn important information in the input sequence, which shows higher accuracy compared to the traditional automatic speech recognition (ASR) model. The Conformer model adds a convolution module to the Transformer's encoder, which increases the ability to obtain subtle local information and further improves the performance of the model. In this study, the Conformer model and the N-gram language model (LM) are used in combination for Chinese speech recognition, and a good recognition effect is obtained. Experiments on the data sets of AISHELL-1 and aidatatang_200zh show that the character error rate of the Conformer model can be reduced to 5.79% and 5.60%, respectively, which is 5.82% and 2.71% lower than that of the Transformer model. Upon the combination with the N-gram LM, the character error rate can be reduced to the optimal performance of 4.86% and 5.10%, respectively, and the real-time factor (RTF) can reach 0.145 66. When the test signal-to-noise ratio is reduced to 20 dB, the character error rate of the model drops to 8.58%, which indicates the anti-noise ability of the model.

Key words: speech recognition; Transformer; language model (LM); Conformer; deep learning

随着科技的飞速发展, 语音识别技术已经成为了智能设备的标配, 这项技术贯穿了多门学科理论, 包含了模式识别、电子技术、数理统计、信号处理、计算机科学、物理声学、生理科学和语言学等. 由于语音

① 基金项目: 山东省重大科技创新工程 (2019JZZZY010120); 山东省重点研发计划 (2019GSF111054)
收稿时间: 2021-10-28; 修改时间: 2021-11-29, 2021-12-13; 采用时间: 2021-12-21; csa 在线出版时间: 2022-05-31

交互提供了更自然、更便利、更高效的沟通形式, 语音必定将成为未来最主要的人机互动接口之一.

在 20 世纪 50 年代, 贝尔实验室就开始基于简单的孤立词语音识别技术的研究[1]. 1968 年, 苏联科学家 Vintsyuk 提出采用动态规划的算法实现动态时间规整 (dynamic time warping, DTW)[2,3], 一度成为当时语音识别的主流技术. 后来模式识别、动态规划算法和线性预测编码这 3 种技术被引入到语音识别中, 成功的使得孤立词语音识别系统从理论上得以完善, 并且可以达到实用化的水平[4,5]. 进入 80 年代后, 基于隐马尔科夫模型 (hidden Markov model, HMM)[6,7] 的声学建模和基于 N-gram 的语言模型在语音识别中得到运用[8,9], 这时期语音识别开始从孤立词识别系统向大量词汇连续语音识别系统发展. 后来又结合高斯混合模型 (Gaussian mixed model, GMM), 形成基于高斯混合模型-隐马尔可夫模型 (Gaussian mixed model-hidden Markov model, GMM-HMM)[10] 的语音识别框架, 使基于 HMM 的语音识别模型效果得到提升.

进入 21 世纪后, 深度学习技术不断发展, 在 2011 年, 微软研究院的 Deng 等人以音素状态为建模单位提出了深度神经网络-隐马尔可夫模型 (DNN-HMM) 的识别方法, 用 DNN 模型代替原来的 GMM 模型, 对每一个状态进行建模, 显著降低了错误率[11]. 但 DNN-HMM 语音识别模型的性能还是会受到数据强制分割、对齐、HMM 遗留的多模块独立训练等问题的限制[12].

到 2015 年, 从联结时序分类算法 (connectionist temporal classification, CTC)[13] 引入到语音识别领域后, 端到端技术开始流行. 端到端技术将整个识别网络简化成一个单一的网络结构, 在训练时只需要注意整个系统的输入和输出, 直接将输入音频序列映射到单词或其他字素序列, 大大减少了对语音识别系统构建的难度, 受到越来越多研究人员的欢迎[14–16].

近几年, 研究人员注意到具有自注意力机制的深度神经网络模型 “Transformer”[17], 在机器翻译、计算机视觉等领域中展现出强劲识别的性能. Dong 等人首次将 Transformer 模型引入到语音识别领域中来, 使得 Transformer 能够完成语音识别任务[18]. Transformer 在提取长序列依赖的时候更有效, 但是提取局部细微特征的能力较弱, 而卷积则更擅长提取局部特征[19–21]. Conformer 模型[22] 将卷积模块加入到 Transformer 模型的编码器部分, 达到增强识别效果的目的. Transformer 模型在推理过程中无需使用语言模型即可获得不错的识别效果, 但所得文本从语言学角度上看质量较差, 结合语言模型之后将得到不错的效果. 本文将 Conformer 模型所搭建的语音识别系统在数据集 AISHELL-1 和 aidatatang_200zh 上与 Transformer 模型作比较, 并且增加语言模型[23,24] 后比较了语音识别系统识别性能以及实时率的差异, 并且在不同程度的噪声数据中测试了识别的准确率.

1 Conformer 模型结构

本文所使用的 Conformer 结构是在 Transformer 模型编码器的基础上增加卷积模块, 构成 Conformer 编码器. 结构如图 1 所示, Conformer 编码器由多个 Conformer 块堆叠而成[22].

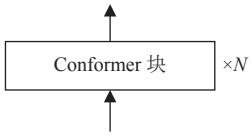

图 1 Conformer 编码器

1.1 Conformer 块

Conformer 模型核心就是编码器中的 Conformer 块, 其结构如图 2 所示, 由 Layer Norm 模块、前馈层、卷积层和多头注意力层组成. 在前馈层、卷积层和多头注意力层上都有残差结构, 这里残差结构的引入是为了便于卷积网络的训练[25]. 同时卷积模块和多头注意力模块相连起到效果增强的作用.

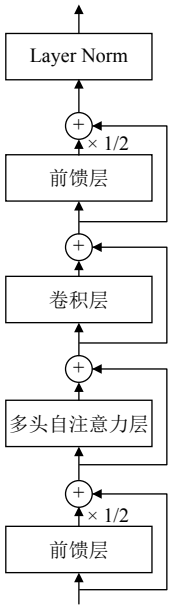

图 2 Conformer 块结构

1.2 多头自注意力层

在多头自注意力模块中, 其结构如图 3 所示, 使用了残差结构和 Dropout 来帮助训练更深层次的网络, 防止多头注意力层向量丢失重要信息[26].

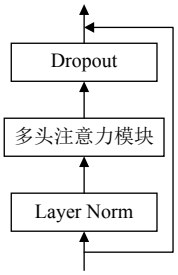

图 3 多头自注意力模块

多头注意力模块中的注意力机制从输入的大量信息中选择关键信息加以处理. 使用信息提取的方法将维度为d_m的输入映射到一组查询Q、键K和值V的矢量输出, 其中查询Q和键K的维度是d_K, 值V的维度是d_V. 然后再利用*Softmax*函数来获得值的权重, 最后返回值的加权总和Z. 计算公式如式 (1) 所示:

$$Z = Attention(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_K}}\right)\cdot V \quad (1)$$

其中, 对QK^{T}相乘结果进行必要的缩放, 来避免值过大导致*Softmax*函数梯度很小难以优化.

多头注意力机制是将h个不同线性变换对Q、K和V进行投影, 最后将不同注意力输出结果拼接起来. 如式 (2)–式 (3) 所示, 多头注意力层输出是将各个注意力头的输出乘以权重矩阵来计算.

$$MultiHead(Q,K,V) = Concat(head_1,\cdots,head_n)W^O \quad (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

其中, W表示线性变换的参数, $head_i$表示第i个注意力头. 多头注意力模块使用了相对正弦位置编码, 这种相对位置编码使自注意模块对不同的输入长度有更好的泛化能力, 并且可使编码器对语音的输入有更好的鲁棒性[27].

1.3 卷积层

Conformer 块结构中的卷积模块如图 4 所示, 由 Layer Norm、Batch Norm、Pointwise 卷积、Depthwise 卷积、GLU 激活层和 ReLU 激活层所组成. 整体运用了残差结构, 增强了梯度的传播, 防止梯度消失[25].

在卷积模块中使用深度可分离卷积, 深度可分离卷积由 Pointwise 卷积和 Depthwise 卷积组成, 它将普通的卷积操作分解为两个过程, 这么做可以用较少的参数学习更丰富的特征并且减少了计算量. Pointwise 卷积运算负责将深度卷积的输出按通道投影到新的特征图上; Depthwise 卷积不同于原始卷积, 一个卷积核负责一个通道, 独立地在每个通道上进行空间卷积[28].

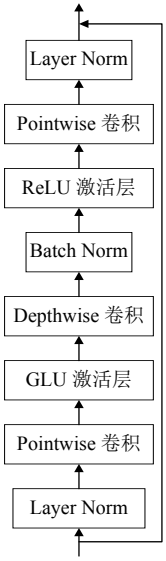

图 4 卷积模块

GLU 激活函数如式 (4) 所示:

$$h_l(X) = (X * W + b) \otimes (X * V + c) \quad (4)$$

其中, W和V是不同的卷积核, b和c是偏置参数, 该函数控制着哪些信息可以传入下一层.

1.4 前馈层

前馈网络 (feed forward network , FFN) 的结构如图 5 所示, 由两个线性层组成, 使用 ReLU 激活函数进行线性变换, 使用 Dropout 层来减少过拟合的发生.

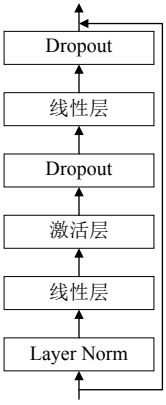

图 5 前馈层结构

如式 (5) 所示, 前馈层目的是为了更新注意力层输出向量的每个状态信息. 其中W表示权重, b表示偏差, x表示输入:

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{5}$$

然后在经过 Layer Norm 层归一化重新定位, 对编码器的深度网络进行平滑优化[29]. 同时控制输入向量长度的动态变化, 防止神经网络层的参数变化导致输入的分布产生较大差异.

2 语言模型

语言模型用于评估文本序列是否符合人类语言使用习惯, 是传统语音识别系统中不可或缺的一部分. 语言模型可以基于语法规则, 也可以基于统计方法. 基于语法规则的语言模型来源于语言学家掌握的语言学领域知识. 而基于统计方法的语言模型, 通过对大量文本语料进行处理, 获得给定词序列出现的概率分布, 以客观描述词与词之间组合的可能性, 适合于处理大规模真实文本.

统计语言模型的目标是计算给定词序列$w_1,\cdots,w_{t-1},w_t$的组合概率, 如式 (6) 所示:

$$\begin{aligned}P(w)&=P(w_1w_2\cdots w_{t-1}w_t)\\&=P(w_1)P(w_2|w_1)\cdots P(w_t|w_1w_2\cdots w_{t-1})\end{aligned} \tag{6}$$

其中, 条件概率$P(w_1),P(w_2|w_1),\cdots,P(w_t|w_1w_2\cdots w_{t-1})$就是语言模型, 计算所有这些概率值的复杂度较高, 特别是长句子的计算量很大, 因此一般采用最多n个词组合的 N-gram 模型. 语言模型的训练需要足够规模的语料数据, 数据越多统计到的词的关系就越多, 概率的区分性也就越明显, 符合语法规范的句子也就越多.

但是, 纯端到端的模型并没有结合语言模型, 在结合语言模型之后会更好地利用中文语言特性得到更加准确的预测结果. 而 N-gram 语言模型有着成熟完备的训练工具, 语料或多或少都可以进行训练并且训练速度也很快, 因此本实验采用 N-gram 语言模型[9].

2.1 N-gram 语言模型

N-gram 是语音识别中最常用到的语言模型. N-gram 指文本中连续出现的n个词语, 基本原理是基于马尔可夫假设, 在训练语料数据中, 通过极大似然估计的方法得到下一个词语出现的n个概率分布进而来推断语句结构.

当n为 1 时称为一元模型, 表示为式 (7):

$$p(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}p(w_i) \tag{7}$$

当n为 2 时称为二元模型, 表示为式 (8):

$$p(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}p(w_i\mid w_{i-1}) \tag{8}$$

当n为 3 时称为三元模型, 表示为式 (9):

$$p=(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}p(w_i\mid w_{i-2}w_{i-1}) \tag{9}$$

多元模型 N-gram 可以表示为式 (10):

$$p(w_1\mid w_{i-n-1},\cdots,w_{i-1})=\frac{C(w_{i-n-1},\cdots,w_i)}{C(w_{i-n-1},\cdots,w_{i-1})} \tag{10}$$

其中, m表示训练语料库中的总字数, $C(w_1,\cdots,w_i)$表示计算$w_1,\cdots,w_i$在训练语料中出现的次数. 一元模型与多元模型相比, 一元模型对句子的约束最小, 其中的竞争最多. 而多元模型对句子有更好的约束能力, 解码效果更好. 但是相应的n越大, 语言模型就越大, 解码速度也就越慢. N-gram 预测的词概率值依赖于前$n-1$个词, 而更长距离的上下文依赖被忽略.

2.2 困惑度和平滑技术

目前主要使用困惑度进行对比来确定语言模型的好坏, 这种指标比较客观. 给定句子S, 其包含词序列$w_1,w_2,\cdots,w_T$, T表示句子的长度, 则其困惑度可以由式 (11) 表示为:

$$PPL(w)=P(w_1w_2\cdots w_T)^{-\frac{1}{T}} \tag{11}$$

困惑度简称为PPL, PPL越小, 句子S出现的概率就越高, 表明语言模型越好, 因此语言模型优化的目标就是最小化困惑度.

语言模型的概率需要通过大量的文本语料来估计, 采用最大似然算法. 但是在统计的预料中数量有限, 因此会存在数据稀疏的情况, 这会导致零概率或估计不准的问题, 因此对预料中未出现或少量出现的词序列, 需要采用平滑技术进行间接预测.

平滑技术主要有 3 种, 有折扣法、插值法和回退法[30]. 折扣法是降低概率不为 0 项的概率, 从已有的观测值概率调配一些给未观测值的概率来提高概率为 0 项的概率, 但没有考虑低阶模型和高阶模型间的关系故不单独使用; 插值法是将高阶模型和低阶模型做线性组合, 充分利用高阶和低阶语言模型, 把高阶的概率信息分配给低阶的模型; 回退法是基于低阶模型估计未观察到的高阶模型.

3 构建语音识别系统

端到端语音识别系统, 不同于传统方法将语音识别任务分解为声学模型、字典和语言模型多个子任务,

而是经过一个复杂网络直接产生对应的语言文本, 并且在不使用语言模型的情况下就能进行语音识别的工作, 实现从输入语音到输出文本的转换[31].

结构如图 6 所示, 编码器部分负责将语音输入序列映射到特征序列, 生成指定长度的向量. 解码器部分对最终的识别结果进行解码, 根据语义向量生成指定的序列.

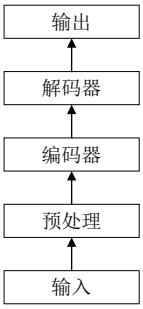

图 6 端到端语音识别系统

预处理模块就是对初始输入进行处理, 如图 7 所示, 该结构是由数据增强层、池化层、线性层和 Dropout 所组成.

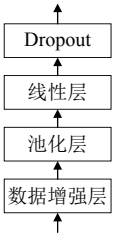

图 7 预处理模块

数据增强层通过使用 SpecAugment[32,33] 方法在 log 梅尔声谱层面上进行数据增强, 可以将模型的过拟合问题转化为欠拟合问题, 以便通过大网络和长时训练策略来缓解欠拟合问题, 提升语音识别效果. 池化层处理输入, 较好地保留了低层次输入, 在保留了编码器的表示能力和模型整体精度的同时显著降低了计算量.

线性层又称为全连接层, 其每个神经元与上一个层所有神经元相连, 实现对前一层的线性组合或线性变换. Dropout 对于神经网络单元按照一定的概率将其暂时从网络中丢弃, 有效地减轻过拟合的发生, 一定程度上达到了正则化的效果.

3.1 端到端结构

端到端模型结构如图 8 所示, 该结构编码器部分为 Conformer 的编码器, 由 12 个 Conformer 块堆叠而成, 解码器部分由 CTC 解码器构成.

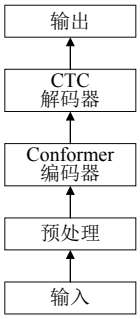

图 8 端到端语音识别系统结构

输入数据经过预处理后进入 Conformer 编码器, CTC 解码器由线性层组成, 将编码器的输出转化为 CTC 激活后解码输出, 解码算法为 CTC Prefix Beam Search[34–36].

CTC 网络的输出形式为$T \times C$, 其中, T 表示时间长度, C表示字符类别数, CTC Prefix Beam Search 算法就是模型读入一帧的数据, 然后给出当下各种字符的概率, 然后利用这一层的概率展开搜索, 取搜索空间中最优的k条路经的前缀, 并把这些前缀挨个输入到模型中, 同时把相同的前缀路径合并, 不断重复最终得到最优解.

3.2 结合语言模型的端到端结构

结合语言模型后的模型结构, 如图 9 所示. 编码器部分由 12 个 Conformer 块组成, 解码器部分为先经过 CTC WFST search 打分后再由 Attention 解码器重新打分得到最终结果[14,37]. 在结合语言模型的结构中, CTC WFST search 是该结构的核心, 该步骤包含了构建解码图和解码器两部分.

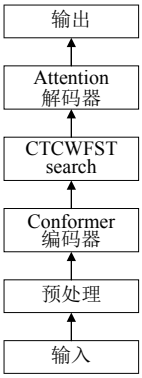

图 9 结合语言模型的结构

解码图用 TLG 来表示, 即将 T、L 和 G 各层次信息组合到一张图中, 其中 T 表示建模单元, L 表示词典, G 表示语言模型. 以端到端模型训练的中文汉字作为建模单元 T, 词典 L 则是由词语或句子拆分成建模单

元而构成, 语言模型 G 是由 N-gram 语言模型转换为加权有限状态转换器 (weighted finite-state transducer, WFST) 的形式表示[38,39]. WFST 通常用来描述状态之间的转移信息, 能够将语言模型直接表示成图的形式, 语言模型概率经处理后作为图中的权重. 当图构建完成之后, 语言模型的概率就成了图权重的一部分, 解码时直接使用图的权重而不用去查询语言模型, 它实现了输入序列到输出序列的转换.

解码器部分采用的是 Viterbi 解码, 根据输入寻求最佳状态序列. 解码过程是逐帧推进, 结合转移弧上的权重, 得到每个时刻扩展路径的累计代价, 然后对比指向同一个状态的不同路径的累计代价, 选择值更小的路径并更新状态信息, 直到 Viterbi 解码最后一帧然后回溯路径, 得到最优路径. 对得到的信息再进行 Attention 解码重打分, Attention 解码器使用 Transformer 结构的解码器部分, 通过使用注意力机制最终输出最合适的结果[37,40,41].

4 实验

4.1 实验数据

实验所用到的语音数据由两部分组成, 一部分来自于北京希尔贝壳科技有限公司出版的中文语声数据集 AISHELL-1, 其包含 178 h 来自 400 个说话人的普通话声频和相应文本信息. AISHELL-1 中的声频数据重采样为 16 kHz, 16 位的 WAV 格式. 开发人员将数据集分为 3 个部分: 训练集、验证集和测试集. 训练集包含来自 340 个说话者的 120 098 个发音和大约 140 h 的普通话语声数据; 验证集包含来自 40 个说话者的 14 326 个语句; 测试集包含来自 20 个说话者的 7 176 个语句. 对于每个说话者, 大约发布了 360 个语句 (大约 26 min 的语声).

另一部分来自于由北京数据堂科技有限公司开发的中文普通话语音语料库 aidatatang_200zh, 语料库包含 200 h 的声学数据, 主要是移动记录数据, 邀请了来自中国不同口音地区的 600 名演讲者参与录音, 每个句子的转录准确率大于 98%, 数据文件中保留了语音数据编码和说话人信息等详细信息.

4.2 实验配置

实验所用的机器操作系统为 Ubuntu 20.04.2LTS, CPU 为 Intel Xeon Silver 4210 , 128 GB 内存, GPU 为 3 块 RTX2080 SUPER (6 GB) 显卡, 共 18 GB 显存.

SpecAugment 使用了 2 个最大频率掩码和 2 个最大时间掩码以缓解过拟合问题. 在编码器的前端使用两个核大小为 3×3、步幅为 2 的卷积子采样层. 编码器中使用 12 个 Conformer 块, 注意力头数设置为 4, 学习率设置为 0.002, batch size 设置为 8, epoch 设置为 120, beam size 设置为 10. Attention 解码器中解码器个数为 6 个, 语言模型使用三元语法模型, 即 N-gram 语言模型中的 N 为 3[32,37].

实验中输入特征是 80 维梅尔滤波器组特征即 Fbank 特征, 将语音通过预加重、分帧、加窗、傅里叶变换、功率谱以及滤波器组有序计算. 设置窗长为 20 ms, 帧移为 10 ms.

训练使用 CTC loss 与 Attention loss 联合优化训练, 这样设置的目的是避免 CTC 对齐关系过于随机还能加快训练的收敛速度, 并且可以使训练过程更加稳定, 从而取得更好的识别结果.

$$L(x,y)=\lambda L_{\mathrm{CTC}}(x,y)+(1-\lambda)L_{\mathrm{ATT}}(x,y) \tag{12}$$

训练所使用的组合损失如式 (12) 所示, x表示声学特征, y为对应标注, $L_{\mathrm{CTC}}(x,y)$表示 CTC loss, $L_{\mathrm{ATT}}(x,y)$表示 Attention loss, λ表示平衡 CTC loss 和 Attention loss 的系数[32,42,43].

本实验基于 Kaldi[44]、Espnet (end-to-end speech processing toolkit) 工具包[45] 和 WeNet[37] 语音识别工具包来进行. Kaldi 是著名的开源语音识别工具, 这套工具提供了目前工业界最常用的模型训练工具, 它使用 WFST 来实现解码算法, 其主要的代码是 C++编写, 在此之上使用 bash 和 Python 脚本做了一些工具. Espnet 工具箱融合了 Kaldi 的数据处理和特征提取, 同时借助 PyTorch 和 Chainer, 使用 Python 实现了许多端到端模型. WeNet 是出门问问语音团队联合西工大语音实验室开源的一款语音识别工具包, 模型训练完全基于 PyTorch 生态, 结构类似于 Kaldi 但并不依赖于 Kaldi 等安装复杂的工具.

4.3 评价标准

本文在数据集 AISHELL-1 和数据集 aidatatang_200zh 上评价实验结果, 采用字错率 (character error rate, CER) 作为评价指标. 字错率即为了使识别出来的词序列和标准的词序列之间保持一致, 需要进行替换、删除或者插入某些词, 这些插入I、替换S和删除D的词的总个数, 除以标准的词序列中词的总个数的百分比, 即如式 (13) 所示:

$$CER=\frac{S+D+I}{N} \tag{13}$$

4.4 实验结果

在数据集 AISHELL-1 和 aidatatang_200zh 上, 不添加语言模型的情况下, 实验结果如表 1, 以 Conformer 模型所搭建的语音识别系统与 Transformer模型做对比, 可以看出在相同的数据集上训练 Conformer 模型较 Transformer 模型具有更低的字错率. 在 AISHELL-l 数据集上 Conformer 模型要比 Transformer模型字错率低 5.82%, 在 aidatatang_200zh 数据集上 Conformer 模型比 Transformer 模型字错率低 2.71%.

表 1 在不同数据集上不同模型的字错率 (%)

语音识别系统	AISHELL-1	aidatatang_200zh
Transformer	11.61	8.31
Conformer	5.79	5.60

添加语言模型之后, 在相同数据集上使用文中识别方法的结果如表 2, 不难看出在 AISHELL-1 数据集上 Conformer 模型在结合语言模型之后比 Transformer 模型结合语言模型的字错率低 3.23%, 在 aidatatang_200zh 数据集上结合语言模型的 Conformer模型比结合语言模型的 Transformer 模型字错率低 1.69%.

表 2 结合语言模型使用不同模型的字错率 (%)

语音识别系统	AISHELL-1	aidatatang_200zh
Transformer	8.09	6.79
Conformer	4.86	5.10

经以上实验表明, 在添加语言模型后 Conformer 模型和 Transformer 模型在两个不同的数据集上准确率均得到了进一步提升, 并且 Conformer 模型在添加语言模型之后识别效果最佳.

语音识别的实时率用来度量语音识别系统识别音频速度的值, 表示处理单位时长语音数据所需要的时间, 值越小表示处理语音的效率越高. 经测试结果如表 3 所示, 在不结合语言模型时 Transformer 模型的实时率比 Conformer 模型低 0.061 02, 在结合语言模型之后 Transformer 模型的实时率比 Conformer 模型低 0.0344, 可以看出 Transformer 模型的实时率比 Conformer 模型的实时率稍好, 并且在结合语言模型之后两模型识别的实时率也均会发生升高, 但仍能在语音识别时达到不错的识别效率.

目前较新的语音识别模型有 RNN-Transducer、Conformer-Transducer[45,46], 以在 AISHELL-1 数据集上测试的结果为基准, 与结合语言模型的 Conformer 模型作比较, 其结果如表 3 所示,

由表 4 可以看出, 结合语言模型的 Conformer 模型较 RNN-Transducer 和 Conformer-Transducer 模型相比, 字错率分别下降了了 2.34% 和 0.14%. 可以看出该模型在性能上有一定的优势.

表 3 语音识别的实时率

语音识别系统	RTF
Transformer	0.067 33
Conformer	0.128 35
Transformer+LM	0.111 26
Conformer+LM	0.145 66

表 4 与目前较新的模型比较字错率 (%)

语音识别模型	字错率
RNN-Transducer	7.20
Conformer-Transducer	5.00
Conformer+LM	4.86

测试结合语言模型的 Conformer 模型在噪声环境的性能, 在 AISHELL-1 数据集上加入不同比例的白噪声分别构成信噪比为 10 dB、20 dB、40 dB、60 dB 和 80 dB 的噪声数据. 测试结果如表 5 所示, 在测试信噪比为 80 dB 和 60 dB 含噪声数据时的性能和与使用纯净音频时的性能十分接近. 随着噪声强度的增加, 在测试信噪分别为 40 dB 和 20 dB 时, 音频质量接近日常生活环境, 此时识别的准确率有所下降. 信噪比为 10 dB 时语音数据声音嘈杂, 对模型的识别产生较大影响, 此时字错率升高. 由此可以看出噪声会对模型的性能产生影响, 随着噪声的增强, 模型识别的准确率有所下降.

表 5 比较在不同噪声环境下的字错率

信噪比 (dB)	字错率 (%)
10	27.30
20	8.58
40	5.08
60	4.96
80	4.96
纯净音频	4.86

5 结束语

本次实验通过比较不同模型的字错率, 可以看出由 Conformer 模型所搭建的中文语音识别系统较 Transformer 模型有更好的性能, 并且语言模型的添加对端到端语音识别系统识别准确的增加有着重要的作用. 模型识别语音的实时率小于 0.2, 在进行语音识别时可以感受到细微的延迟并不会影响整体的效果. 并且通过在含有不同程度噪声数据上测试的结果, 可以看出不同程度的噪声均会对模型的性能产生一定的影

响. 由于实验中所用于训练的语音数据是在安静的条件下录制的, 语音质量比较高, 这相较于模型在实际使用中所输入的语音数据过于完美, 并且实验所用的数据量不足无法涉及到现实中的各个生活场景, 因此后续考虑扩充实验数据量以提升模型的性能及鲁棒性, 使该模型能够在更多环境下使用.

参考文献

1 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]

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

3 Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(1): 43–49. [doi: 10.1109/TASSP.1978.1163055]

4 Buzo A, Gray A, Gray R, *et al*. Speech coding based upon vector quantization. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(5): 562–574. [doi: 10.1109/TASSP.1980.1163445]

5 Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357–366. [doi: 10.1109/TASSP.1980.1163420]

6 Lee KF, Hon HW, Hwang MY, *et al*. The SPHINX speech recognition system. International Conference on Acoustics, Speech, and Signal Processing. Glasgow: IEEE, 1989: 445–448.

7 Juang BH, Rabiner LR. Hidden Markov models for speech recognition. Technometrics, 1991, 33(3): 251–272. [doi: 10.1080/00401706.1991.10484833]

8 Huang XD, Baker J, Reddy R. A historical perspective of speech recognition. Communications of the ACM, 2014, 57(1): 94–103. [doi: 10.1145/2500887]

9 Cavnar WB, Trenkle JM. N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas. 1994. 161175.

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

11 Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: An overview. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 8599–8603.

12 Wang D, Wang XD, Lv SH. An overview of end-to-end automatic speech recognition. Symmetry, 2019, 11(8): 1018. [doi: 10.3390/sym11081018]

13 Graves A, Fernández S, Gomez F, *et al*. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh: ACM, 2006. 369–376.

14 Miao YJ, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Scottsdale: IEEE, 2015. 167–174.

15 Lu L, Zhang XX, Renais S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016. 5060–5064.

16 Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. Proceedings of the 31st International Conference on Machine Learning. Beijing: ACM, 2014. II-1764–II-1772.

17 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

18 Dong LH, Xu S, Xu B. Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5884–5888.

19 Bie A, Venkitesh B, Monteiro J, *et al*. A simplified fully quantized transformer for end-to-end speech recognition. arXiv: 1911.03604, 2019.

20 刘长征, 张磊. 语音识别中卷积神经网络优化算法. 哈尔滨理工大学学报, 2016, 21(3): 34–38.

21 杨洋, 汪毓铎. 基于改进卷积神经网络算法的语音识别. 应用声学, 2018, 37(6): 940–946. [doi: 10.11684/j.issn.1000-310X.2018.06.016]

22 Gulati A, Qin J, Chiu CC, *et al*. Conformer: Convolution-augmented transformer for speech recognition. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, 2020. 5036–5040.

23 Graovac J, Mladenović M, Tanasijević I. NgramSPD: Exploring optimal n-gram model for sentiment polarity detection in different languages. Intelligent Data Analysis, 2019, 23(2): 279–296. [doi: 10.3233/IDA-183879]

24 Pibiri GE, Venturini R. Handling massive *N*-gram datasets efficiently. ACM Transactions on Information Systems, 2019, 37(2): 25.

25 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

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

27 Bello I, Zoph B, Le Q, *et al*. Attention augmented convolutional networks. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3285–3294.

28 Kriman S, Beliaev S, Ginsburg B, *et al*. Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions. Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6124–6128.

29 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning. Lille: ACM, 2015. 448–456.

30 徐望, 王炳锡. N-gram 语言模型中的插值平滑技术研究. 信息工程大学学报, 2002, 3(4): 13–15. [doi: 10.3969/j.issn.1671-0673.2002.04.004]

31 杨鸿武, 周刚. 基于改进混合 CTC/attention 架构的端到端普通话语音识别. 西北师范大学学报 (自然科学版), 2019, 55(3): 48–53.

32 Park DS, Chan W, Zhang Y, *et al*. SpecAugment: A simple data augmentation method for automatic speech recognition. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, 2019. 2613–2617.

33 Park DS, Zhang Y, Chiu CC, *et al*. Specaugment on large scale datasets. Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6879–6883.

34 杨德举, 马良荔, 谭琳珊, 等. 基于门控卷积网络与 CTC 的端到端语音识别. 计算机工程与设计, 2020, 41(9): 2650–2654.

35 杨威, 胡燕. 混合 CTC/attention 架构端到端带口音普通话识别. 计算机应用研究, 2021, 38(3): 755–759.

36 Hannun AY, Maas AL, Jurafsky D, *et al*. First-pass large vocabulary continuous speech recognition using Bi-directional recurrent dnns. arXiv: 1408.2873, 2014.

37 Zhang BB, Wu D, Yang C, *et al*. WeNet: Production first and production ready end-to-end speech recognition toolkit. arXiv: 2102.01547, 2021.

38 Mohri M, Pereira F, Riley M. Speech recognition with weighted finite-state transducers. Benesty J, Sondhi MM, Huang YA. Springer Handbook of Speech Processing. Berlin, Heidelberg: Springer, 2008. 559–584.

39 Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition. Computer Speech & Language, 2002, 16(1): 69–88.

40 张晓旭, 马志强, 刘志强, 等. Transformer 在语音识别任务中的研究现状与展望. 计算机科学与探索, 2021, 15(9): 1578–1594. [doi: 10.3778/j.issn.1673-9418.2103020]

41 胡章芳,蹇芳,唐珊珊. DFSMN-T: 结合强语言模型 Transformer 的中文语音识别. 计算机工程与应用, 2022, 58(9): 187–194.

42 Watanabe S, Hori T, Kim S, *et al*. Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240–1253. [doi: 10.1109/JSTSP.2017.2763455]

43 Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 4835–4839.

44 Povey D, Ghoshal A, Boulianne G, *et al*. The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding Workshop. Hawaii: IEEE, 2011.

45 Watanabe S, Hori T, Karita S, *et al*. Espnet: End-to-end speech processing toolkit. 19th Annual Conference of the International Speech Communication Association. Hyderabad, 2018. 2207–2211.

46 Rao K, Sak H, Prabhavalkar R. Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Okinawa: IEEE, 2017. 193–199.

(校对责编: 牛欣悦)