基于ADTDNN的低资源语音识别方法研究

顾龙昊[1,2],黄连丽[1],周 奎[2],张子越[1,2]

(1. 湖北汽车工业学院 电气与信息工程学院;
2. 湖北汽车工业学院 汽车工程师学院 Sharing-X重点联合实验室,湖北 十堰 442002)

摘 要: 为解决低资源条件下由于训练数据不足导致识别精度降低、泛化能力较差的问题,提出一种语音识别方法。该方法利用卷积池化提取特征信息,将Attention机制与DTDNN融合成为ADTDNN,以提升低资源环境下模型捕捉序列中关键信息的能力;采用链接时序分类简化模型的识别流程;使用Transformer作为语言模型。在Aishell-1数据集上的实验结果表明,低资源环境下基于ADTDNN的语音识别模型与LAS、Transformer等主流端到端模型相比,字错误率分别降低了3.7%和1.0%。

关键词: 语音识别;时延神经网络;Transformer;数据增强;低资源
DOI: 10.11907/rjdk.232097 **开放科学(资源服务)标识码(OSID):**
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0076-06

Research on Low-Resource Speech Recognition Based on ADTDNN

GU Longhao[1,2], HUANG Lianli[1], ZHOU Kui[2], ZHANG Ziyue[1,2]

(1. *Department of Electrical and Information Engineering, Hubei University of Automotive Technology*;
2. *Sharing-X Key Joint Laboratory, Institute of Automotive Engineers, Hubei University of Automotive Technology, Shiyan* 442002, *China*)

Abstract: A speech recognition approach has been proposed to address the problem of reduced recognition accuracy and poorer generalization performance due to insufficient training data in low-resource conditions. This method leverages convolutional neural networks to extract feature information. It combines the attention mechanism with delayed time-delay neural networks, referred to as ADTDNN, enhancing the model's ability to capture key information in sequences within low-resource environments. The approach employs linking temporal classification to streamline the recognition process of the model. Additionally, a Transformer is utilized as the language model. Experimental results on the Aishell-1 dataset demonstrate that the ADTDNN-based speech recognition model in low-resource settings reduces word error rates by 3.7% and 1% compared to mainstream end-to-end models like LAS and Transformer, respectively.

Key Words: speech recognition; time delay neural networks; Transformer; data enhancement; low resource

0 引言

自动语音识别技术的目的在于使计算机能够识别人类的声音信息,并将其转换为文字。在深度学习热潮之前,统计模型GMM-HMM(Gaussian Mixture Model-Hidden Markov Model)一直占据主导地位[1]。其使用的主要技术为:①采用隐马尔可夫模型(Hidden Markov Model,HMM)对语音状态的转移概率进行建模;②采用高斯混合模型(Gaussian Mixture Model,GMM)对语音状态的观察值概率进行建模。

2009年,Hinton等[2]使用深度神经网络完成了声学模型的构建,取得了重大突破,至此基于深度学习的语音识别模型逐渐成为主流。当前基于深度学习的语音识别模型主要分为深度神经网络-HMM(Deep Neural Network-HMM,DNN-HMM)模型和端到端模型两类[3]。DNN-HMM模型构建较为复杂,需要实现数据帧与状态的对齐等。端到端语音识别模型构建相对简单,只需要音频及其对应的文本标签即可直接将语音转换为文本。目前端到端语音识别模型主要包括基于注意力机制的编解码模型

收稿日期: 2023-10-27
扫描二维码阅读全文:
基金项目: 湖北省重点研发计划项目(2023BAB169);湖北省武汉市科技重大专项(2022013702025184)
作者简介: 顾龙昊(2000-),男,湖北汽车工业学院电气与信息工程学院硕士研究生,研究方向为语音识别、自然语言处理;黄连丽(1981-),女,硕士,湖北汽车工业学院电气与信息工程学院副教授、硕士生导师,研究方向为智能驾驶。

（Attention based Encoder-Decoder，AED）[4]、循环神经网络转换器（RNN-Transducer，RNN-T）[5]、Transformer[6]、Conformer[7]等。

虽然基于深度学习的语音识别技术相对成熟，但是在低资源环境下缺乏大量标记训练数据来对抗过拟合问题，无法保证识别的准确性[8]。上述问题通常使用数据增强策略来解决，如调整音频的速度或音高、在语音数据中添加噪声、对频谱图进行变换以及使用SpecAugment[9]对频谱图的时域和频域进行屏蔽等。上述方法着重于改变语音输入，而不改变相应的文本标签，并且需要复杂的超参数设置。为此，本文将卷积神经网络（Convolutional Neural Networks，CNN）、Attention机制、密集连接时延神经网络（Densely Connected Time Delay Neural Network，DTDNN）[10]、连接时序分类（Connectionist Temporal Classification，CTC）[11]作为声学模型，将Transformer作为语言模型，同时结合MixSpeech数据增强策略[12]，提出一种基于注意力机制的密集连接时延神经网络（Attention- Densely Connected Time Delay Neural Network，ADTDNN），对低资源环境下的端到端语音识别方法进行研究，以期提高语音识别准确率。

1　理论基础

语音识别系统分为声学模型和语言模型两部分，其根本任务是将语音信号转换为文本信息。从数学角度来看，语音识别的任务为根据输入的音频特征向量O，找到最可能的词序列W。使用公式表示为：

$$\hat{W} = \arg\max P(W|O) = \arg\max \frac{P(W)P(O|W)}{P(O)} \tag{1}$$

$$\hat{W} = \arg\max P(W)P(O|W) \tag{2}$$

式中：$P(O|W)$表示声学模型，使用语音数据进行训练；$P(W)$表示语言模型，使用文本数据进行训练。

1.1　时延神经网络

时延神经网络（Time Delay Neural Network，TDNN）亦称为1维卷积神经网络（1-dimension CNN），由Waibel等[13]提出。相较于传统DNN的全连接，TDNN的上下文具有较强的关联性，其输入不只与当前时刻有关，还与过去、未来的输入相关。TDNN具有多种变体，如因式分解的时延神经网络（Factorized TDNN，TDNNF）[14]、扩展的时延神经网络（Extend TDNN，ETDNN）[15]。TDNNF通过奇异值分解（Singular Value Decomposition，SVD）将权重矩阵分解为两个子矩阵，通过丢弃较小的奇异值来优化网络参数，同时该结构使用了类似ResNet的跳层连接来缓解梯度消失问题，提高模型的泛化能力。ETDNN通过在每个权重矩阵后添加一个前馈网络层来组成交错重复的结构组合。上述两种结构虽然极大提高了传统TDNN的识别精度，但因具有深层次的网络结构而大幅度增加了参数量。DTDNN很好地解决了这个问题，在保证网络结构深度的同时具有更少的网络参数。

1.2　n-gram语言模型

n-gram（n元模型）语言模型用于表示n个词之间的关系，预测的当前词只与n-1个词相关[16]，当n=1、2、3时分别称为一元（unigram）、二元（bigram）、三元模型（trigram）。n取值越大，模型区分性越好，但同时降低了可靠性。权衡考虑区分性与可靠性，以n=3最为合适。n-gram模型可用公式表达为：

$$P(W_n|W_1W_2...W_{n-1}) = \frac{count(W_1W_2...W_n)}{count(W_1W_2...W_{n-1})} \tag{3}$$

式中：W_n表示词序列；$P(W_n|W_1W_2...W_{n-1})$表示在给定$W_1W_2...W_{n-1}$的情况下出现W_n的概率；$count(W_1W_2...W_n)$表示词序列在训练语料中出现的次数。

n-gram通常用作语音识别中的语言模型，但其只依赖前n-1个词，没有较好地上下文建模能力，因此引入上下文关联性强的Transformer作为本文语言模型，以提高上下文建模能力。

2　基于ADTDNN的语音识别模型构建

本文构建的语音识别模型主要包括MixSpeech、声学模型、语言模型3个部分。结构如图1所示。

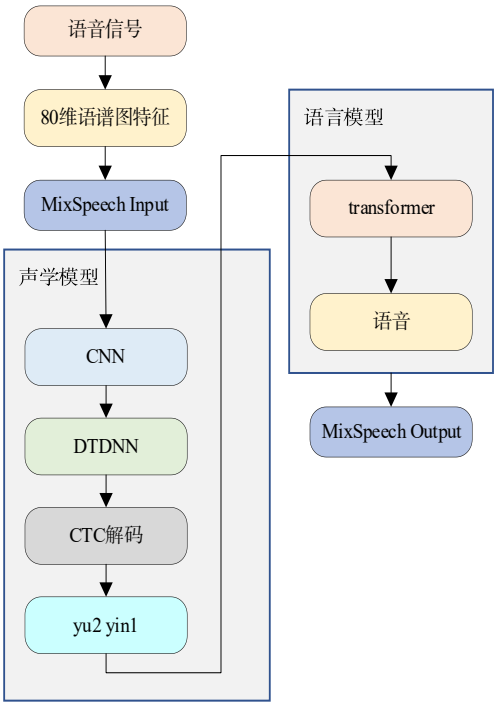

Fig. 1　Speech recognition model structure

图1　语音识别模型结构

2.1　MixSpeech

MixSpeech是一种简单有效的数据增强方法，超参数只有一个组合权重λ。其将两个输入的语音序列逐帧加权相加以得到混合语音序列，通过权重组合系数λ确定混合

比例，同时识别两个文本序列来训练语音识别模型。语音识别模型的两个损失使用与输入相同的权重进行组合，得到混合权重损失。MixSpeech 的结构如图 2 所示，计算公式为：

$$X_{mix} = \lambda X_i + (1 - \lambda) X_j \tag{4}$$

$$Loss_i = Loss(X_{mix}, Y_i) \tag{5}$$

$$Loss_j = Loss(X_{mix}, Y_j) \tag{6}$$

$$Loss_{mix} = \lambda Loss_i + (1 - \lambda) Loss_j \tag{7}$$

式中：X_i 和 X_j 为两个不同的频谱图；X_{mix} 为两个频谱图的权重混合，权重为 λ；y_i、y_j 分别为 X_i 与 X_j 的目标文本序列；$Loss_i$ 为 X_{mix} 及其相应的文本标签 Y_i 计算得到的损失值；$Loss_j$ 的计算方式与 $Loss_i$ 相同；$Loss_{mix}$ 为两个损失函数的权重组合，用于在训练期间对模型进行更新。

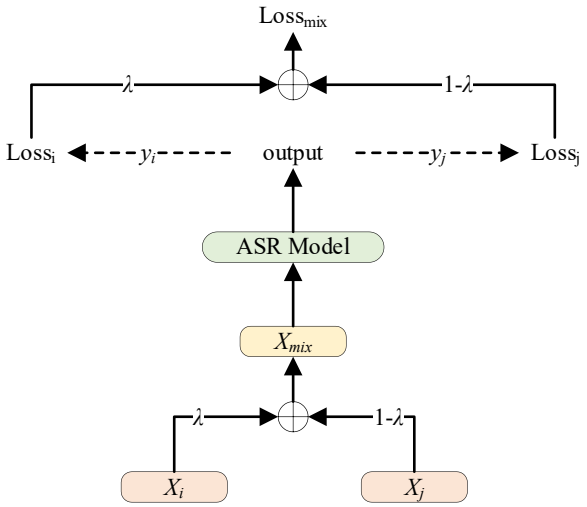

Fig. 2 MixSpeech structure

图 2 MixSpeech 结构

2.2 声学模型

2.2.1 CNN-DTDNN

本文提出的声学模型综合了 CNN 与 DTDNN 网络结构的优点，共包括卷积池化层、DTDNN 块、CTC 解码层 3 个组件。具体结构如图 3 所示。

卷积池化层共 4 层，每层都由两个卷积层和一个池化层组成。该层用于提取语谱图局部特征，并通过最大池化将特征进一步压缩，以减少参数量。

DTDNN 块由多个 DTDNN 层组成，每个 DTDNN 层的结构与 DenseNet 类似，不同的是 DenseNet 中的 CNN 被替换为前馈神经网络（Feedforward Neural Network，FNN）和 TDNN。每个 DTDNN 块的功能不同，第 1 个 DTDNN 块的帧偏移为 1，用于学习本地特征；第 2、3 个 DTDNN 块帧偏移为 3，用于捕获长期依赖关系。DTDNN 块之间使用一个 FNN 来进行连接，用于聚合上下层信息，使模型能够学习到更多过去和未来的信息，从而提高上下文建模能力。

CTC 的作用为在输入序列与输出序列之间建立映射关系，以找到最佳匹配。本文中的 CTC 用于序列解码，输出拼音序列。

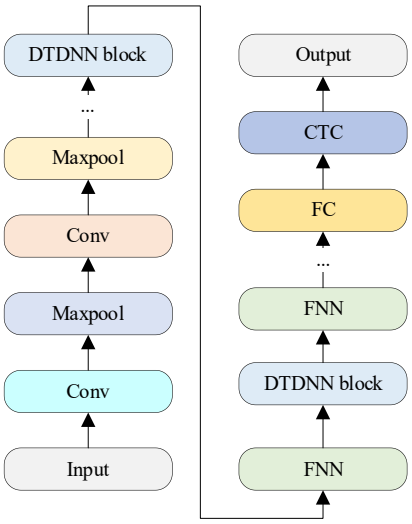

Fig. 3 CNN-DTDNN network structure

图 3 CNN-DTDNN 网络结构

2.2.2 改进 DTDNN

为进一步提升声学模型的性能，本文对 DTDNN 进行改进，将 Attention 与 DTDNN 层进行融合组成 ADTDNN，其中 Attention 可以增强网络处理序列数据的能力，动态计算每个时间步的注意力权重。模型结构如图 4 所示。具体改进方法为：在每个 DTDNN 层后添加一个注意力机制模块，在注意力机制模块中根据输入序列和上下文的关系来计算权重矩阵，最后将当前层的输出与上一层的输出进行加权求和。注意力权重的计算公式为：

$$Q = W_q * y \tag{8}$$

$$K = W_k * y \tag{9}$$

$$V = W_v * y \tag{10}$$

$$O = \frac{soft\max(Q * K^T)}{\sqrt{dk}} * V \tag{11}$$

式中：W_q、W_k、W_v 为随机初始化的权重矩阵；y 表示 DTDNN 模型的隐层输出；dk 表示注意力头的维度。加权求和公式为：

$$out = \mathrm{Re}LU(W * (y + O) + c) \tag{12}$$

式中：W、c 为权重参数。

使用单个 Attention 拟合能力较差，在实际应用中常使用由多个注意力模型并联而成的多头注意力机制。计算公式为：

$$MultiHeadAttention = Concatenate(O_1 O_2 ... O_n) W_0 \tag{13}$$

式中：W_0 表示一个新的随机初始化矩阵。

2.3 Transformer 语言模型

Transformer 是一种完全基于自注意力机制（Self-attention）的网络结构，可以建立输入序列与输出序列之间的对应关系。Transformer 由多个编码器和解码器组成，每个编码器和解码器的输出都会作为下一层编码器和解码器的输入。此外，Transformer 可以扩展到相当深的层数，极大提高了模型的准确性。其结构如图 5 所示。

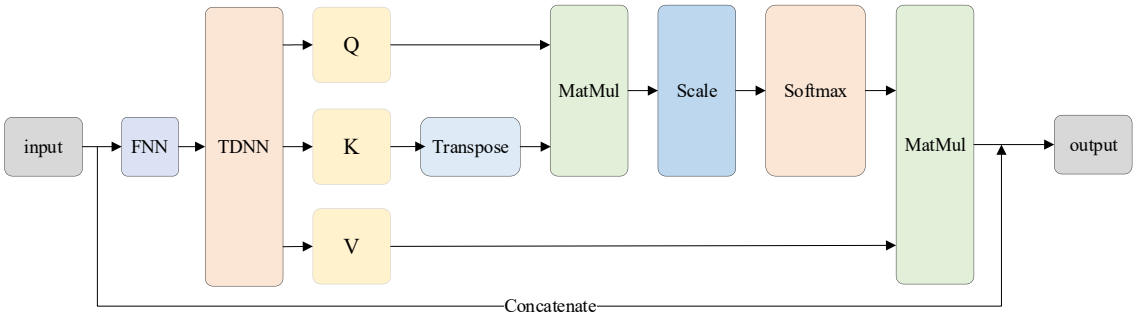

Fig. 4　ADTDNN network structure

图4　ADTDNN网络结构

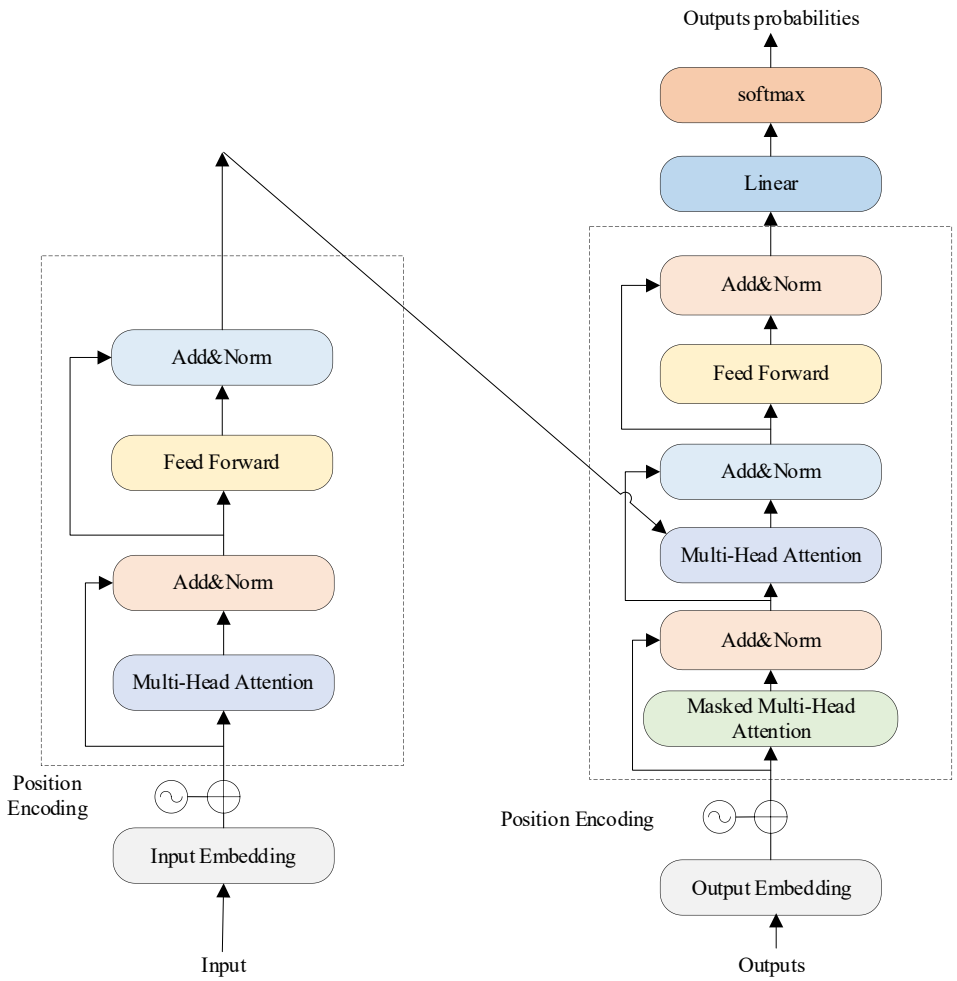

Fig. 5　Transformer network structure

图 5　Transformer网络结构

2.3.1　编码器

编码器的的主要作用为提取音素特征,主要由 Attention、叠加(Add)、层归一化(Layer Norm)和FFN 4个部分组成。每层 Attention都有残差网络进行连接,用于为下层提供更丰富的特征;Layer Norm 的作用为加速模型训练过程,加快收敛速度;FNN由两个线性变换和一个ReLU激活函数组成,用公式表述为:

$$FNN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{14}$$

式中:x表示输入向量;W_n为第n层权重矩阵;b_n为第n层偏置向量;$\max(0, xW_1 + b_1)$为修正单元激活函数,用于保留正值,将负值归零。

2.3.2　解码器

解码器相较于编码器在结构上增加了一个带掩码的多头注意力层(Masked Multi-Head Attention),其利用掩码将下文位置的权重屏蔽,从而使模型在输出时只依赖当前和过去的输入。解码器需要一步一步去解码,最后通过

Softmax层输出词的概率分布。

3 实验方法与结果分析

3.1 实验环境

硬件环境为Intel CoreTM i7 12650H CPU、NVIDIA RTX2080Ti（16GB）GPU、内存为32 GB。操作系统为Ubuntu20.04，Python版本为3.8.0。实验均在PyTorch 2.0.1深度学习框架下进行。

3.2 实验数据集与评价指标

本文实验数据来源于希尔贝壳提供的Aishell-1中文普通话数据集，采样率为16 kHz，时长为120 h。测试集和验证集分别为总数据集的20%和10%，用于不同模型效果比较实验和模型改进前后比较实验。在该数据集中抽取约10%的数据作为低资源数据集，测试集与验证集分布与上述相同，用于数据增强效果比较实验。

使用字错误率（Character Error Rate，CER）作为模型评价指标。计算公式为：

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

式中：N为原字符串长度，S为替换的字符数，D为删除的字符数，I为插入的字符数。

3.3 超参数设置

MixSpeech混合权重λ~Beta(a,a)，其中$a \in (0,\infty)$，本文中Beta分布的a设置为0.5。声学模型的语音信号采样率为16 kHz，每帧长25 ms，对每帧信号加汉明窗（Hamming）进行分析；提取80维FBank语谱图特征作为声学模型输入，batch size设置为8，多头注意力数为8，学习率为1E-4，dropout设置为0.5，使用Adam进行优化。语言模型部分的特征为200维，batch size设置为16，多头注意力数为8，学习率为1E-4，dropout设置为0.5。

3.4 结果分析

3.4.1 不同模型效果比较

对CNN-DTDNN-Transformer、CNN-BiLSTM[17]、DFSMN-3-gram[18]、DFSMN-Transformer[19]、Dual DCNN[20]5种模型以及CNN-DTDNN-Transformer的改进型CNN-ADTDNN-Transformer（本文模型）进行比较实验，结果如表1所示。可以看出，本文提出的基础模型CNN-DTDNN-Transformer在Aishell-1数据集上的准确率相较其他几种模型有较为明显的提升，字错误率降至13.5%，而改进型CNN-ADTDNN-Transformer更是将字错误率降至11.3%。

3.4.2 不同数据增强策略比较

使用不同数据增强策略进行实验以验证本文方法的优势。分别使用MixSpeech、SpecAugment、TTS（Text to Speech）策略进行数据增强，结果如表2所示。可以看出，MixSpeech获得了最低的字错误率。MixSPeech与其他方法相比参数简单，只有一个权重参数λ，其还引入了对比信号，使模型能够更好地识别语音的对应文本，不受其他语音的干扰，因此识别效率高。

Table 1 Effect of different models

表1 不同模型效果比较 %

语音模型	CER
CNN-DTDNN-Transformer	13.5
CNN-BLSTM	19.2
DFSMN-3-gram	15.0
DFSMN-Transformer	13.9
Dual DCNN	15.4
CNN-ADTDNN-Transformer	11.3

Table 2 Comparison of different data enhancement strategies

表2 不同数据增强策略比较 %

数据增强	CER
MixSpeech	21.9
SpecAugment	23.5
TTS	24.1
无数据增强	27.6

3.4.3 低资源环境下不同模型效果比较

对本文模型在低资源环境中的识别效果进行评估，使用LAS（Listen，Attend and Spell）、Transformer作为对照，结果如表3所示[21]。可以看出，本文模型在低资源环境下的识别精度相较Transformer与LAS分别提升了1.0%和3.7%，优势明显。

Table 3 Effect of different models in low resource environments

表3 低资源环境下不同模型效果比较 %

语言模型	CER
CNN-ADTDNN-Transformer	21.9
Transformer	22.9
LAS	26.6

4 结语

本文首先将DTDNN、CNN与CTC相结合提出声学模型CNN-DTDNN-CTC，并使用Transformer作为语言模型，构成端到端的语音识别模型；然后对DTDNN进行改进，使用Attention进行优化，提出CNN-ADTDNN-Transformer模型，进一步提升了识别精度；最后使用MixSpeech对模型进行优化，提升其在低资源环境下的识别精度以及泛化能力。该模型有效利用了DTDNN建模上下文、捕捉序列时间依赖特征的能力以及Attention捕捉关键特征的能力，在Aishell-1数据集上取得了较低的字错误率。

本文模型在低资源环境下具有上下文建模能力强、识别精度高的特点，非常适合于自动驾驶、车路协同、高精度农业智能系统等场景，具备一定的实用意义和推广价值，未来将结合具体的应用场景推进产业化进程。本文模型虽然在性能方面有明显提升，但是由于注意力模块的每个注意力头都有自己的权重矩阵，且计算内容可能会重复，会导致训练时间增加以及数据冗余问题，因此未来将针对模型轻量化以及降维方向继续展开研究。

参考文献:

[1] GALES M,YOUNG S. The application of hidden Markov models in speech recognition[J]. Foundations and Trends® in Signal Processing,2008,1(3):195-304.

[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] DAHL G E,YU D,DENG L,et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(1): 30-42.

[4] 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, 2016: 4945-4949.

[5] KAWAKAMI K. Supervised sequence labelling with recurrent neural networks[D]. Munich:Technical University of Munich, 2008.

[6] VASWANI A,SHAZEER N,PARMAR N, et al. Attention is all you need[C]//31st International Conference on Neural Information Processing Systems,2017:6000-6010.

[7] GULATI A, QIN J, CHIU C C, et al Conformer: convolution-augmented transformer for speech recognition [DB/OL]. https://arxiv. org/pdf/2005. 08100.

[8] YU C C,KANG M,CHEN Y B,et al. Acoustic modeling based on deep learning for low-resource speech recognition: an overview[J]. IEEE Access,2020(8):163829-163843.

[9] PARK D S, CHAN W, ZHANG Y, et al. Specaugment: a simple data augmentation method for automatic speech recognition[DB/OL]. https://arxiv. org/abs/1904. 08779.

[10] YU Y Q, LI W J. Densely connected time delay neural network for speaker verification[C]//Interspeech,2020:921-925.

[11] 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.

[12] MENG L, XU J, TAN X, et al. Mixspeech: data augmentation for low-resource automatic speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing,2021:7008-7012.

[13] WAIBEL A,HANAZAWA T,HINTON G,et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing,1989,37(3):328-339.

[14] POVEY D, CHENG G, WANG Y, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks[C]//Interspeech,2018:3743-3747.

[15] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recognition for multi-speaker conversations using x-vectors[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 5796-5800.

[16] AO J, KO T. Improving attention-based end-to-end ASR by incorporating an N-gram neural network[C]//12th International Symposium on Chinese Spoken Language Processing,2021:1-5.

[17] WANG D, WANG X, LV S. End-to-end mandarin speech recognition combining CNN and BLSTM[J]. Symmetry, 2019, 11(5): 644.

[18] ZHANG S,LEI M,YAN Z,et al. Deep-FSMN for large vocabulary continuous speech recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing,2018:5869-5873.

[19] HU Z F,JIAN F,TANG S S,et al. DFSMN-T: Chinese speech recognition combining strong language model Transformer[J]. Computer Engineering and Applications,2022,58(9):187-194.
胡章芳,蹇芳,唐珊珊,等. DFSMN-T:结合强语言模型Transformer的中文语音识别[J]. 计算机工程与应用,2022,58(9):187-194.

[20] ZHANG Y, LI H Y,XING L,et al. Chinese speech recognition based on dual-path convolutional neural network [J]. Computer Engineering and Design,2023,44(3):880-886.
张昱,李鸿燕,邢璐,等. 基于双路卷积神经网络的中文语音识别[J]. 计算机工程与设计,2023,44(3):880-886.

[21] WILLIAM C,NAVDEEP J, QUOC V L,et al. Listen, attend and spell [DB/OL]. https://arxiv. org/abs/1508. 01211.

(责任编辑:尹晨茹)