多任务学习在中国方言分类中的应用研究

万 苗[1],任 杰[1*],马 苗[1],曹 瑞[2]
(1. 陕西师范大学 计算机科学学院,陕西 西安 710119;
2. 西北大学 信息科学与技术学院,陕西 西安 710127)

摘 要:近年来,随着深度学习技术在语音识别领域的出色表现,基于深度学习的语音识别系统被广泛应用于智能家居、智能客服、会议纪要、实时字幕等多个应用场景。但由于中国民族众多,语言文化差异大、方言多样复杂等特点,给语音识别系统带来了很大的挑战,特别针对短时语音段方言识别任务,已有的中国方言分类系统性能依然较差。针对特征参数梅尔倒谱系数(mel-scale frequency cepstral coefficients,MFCC)进行研究分析,面向中国十种方言数据集构建基于深度学习的方言分类模型。首先,针对 MFCC 构建基于短期记忆网络(long short-term memory,LSTM)的单任务学习模型,准确率可达 79.04%;然后,深入挖掘方言地域特征,提出以方言所在区域为辅助任务的多任务模型,构建基于参数硬共享的多任务学习模型,实验结果显示,分类准确率最高可达 79.96%;最后,针对参数硬共享无法有效挖掘子任务间关联性的问题,首次提出基于参数稀疏共享的多任务学习模型,该模型通过联合训练,自动挖掘子任务间相关性,裁剪多余网络,并进行网络参数共享,实验结果显示,提出的基于 MFCC 特征的参数稀疏共享的多任务分类模型性能最优,分类准确率最高可达 83.59%。

关键词:中国方言分类;多任务学习;神经网络;MFCC;神经网络参数共享
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)04-0109-07
doi:10.3969/j.issn.1673-629X.2022.04.019

Chinese Dialect Classification via Multi-task Learning

WAN Miao[1],REN Jie[1*],MA Miao[1],CAO Rui[2]
(1. School of Computer Science,Shaanxi Normal University,Xi'an 710119,China;
2. School of Information and Technology,Northwest University,Xi'an 710127,China)

Abstract:Recently,with the outstanding performance of deep learning technology in the field of speech recognition,deep learning-based speech recognition systems have been widely used in multiple application scenarios such as smart homes,intelligent customer service,meeting minutes,and real-time translation. However,due to a large number of ethnic groups in China,there are many differences in language and culture,which is a big challenge to the speech recognition system,especially for short-term speech segment dialect recognition tasks,the performance of the existing Chinese dialect classification system is still poor. We research and analyze MFCC and build the dialect classification model based on deep neural network for a dataset of 10 dialects in China. First,we propose a single-task learning model based on the LSTM (long short-term memory) for the three speech features. The single-task model achieves the highest accuracy of 79.04% compared with the other two feature-based models. Second,we study the geographical features of dialects and propose a multi-task model that uses the area of the dialect as an auxiliary task and builds a hard parameter sharing based multi-task learning model. The results show that the performance of this model can achieve up to 79.96%. Finally,due to the hard parameter sharing cannot effectively lean the correlation between sub-tasks,we propose a sparse parameter sharing based multi-task learning model. The model uses joint training to automatically learn the correlation between sub-tasks,prune redundant networks,and share network parameters. Experimental results show that compared with other SOTA method,the multi-task classification model based on MFCC features with sparse parameter sharing performs best,with a classification accuracy of 83.59%.

Key words:Chinese dialect recognition;multi-task learning;neural network;MFCC;parameter sharing of neural network

收稿日期:2021-04-25 **修回日期**:2021-08-26
基金项目:国家自然科学基金(61902229,61872294);陕西省国际科技合作计划项目-一般项目(2020KW-006);中央高校基本科研业务专项资金(GK201803063)
作者简介:万 苗(2001-),女,研究方向为自然语言处理和移动计算;通讯作者:任 杰(1988-),男,博士,讲师,CCF 会员(77097M),研究方向为移动计算和系统性能优化。

0 引 言

语音识别是人机交互的重要组成部分,现如今,基于深度神经学习的语音识别系统日趋成熟,并在导航、翻译、智能家居、车载系统、教学等诸多领域得到广泛应用[1]。但目前由于用户输入语音可能存在口音、方言等特征,导致智能语音系统常常出现无法准确识别的问题[2],进而需要用户矫正口音、重复输入语音指令,严重影响用户使用体验。由此,预先自动判定输入音频语种是提升语音识别系统后端效能的关键步骤。

此外中国历史悠久,地域跨度大,民族繁多,方言种类多样且差异性较大,由于普通话的推广及普及,方言使用频率日趋减少。由此,中共中央办公厅、国务院办公厅印发《关于实施中华优秀传统文化传承发展工程的意见》,提出要“大力推广和规范使用国家通用语言文字,保护传承方言文化”,各地政府也积极建立方言语料库及方言博物馆。进一步,对方言的准确识别及翻译可以有效提升方言使用范围及频率,对方言文化的保护至关重要。

具体的说,方言语种识别是对输入音频源数据的语种种属进行分析判定,进一步根据语种分类结果将音频数据输入到对应的自然语言处理模型进行推理。该文针对方言语种分类进行研究,基于深度学习模型构建方言语种分类模型,面向西北、华北、西南、华中地区的十地方言进行语种分类实验验证。具体来说,主要贡献如下:

(1)抽取 MFCC 特征,构建基于 MFCC 的 LSTM 单任务深度学习模型,进一步,考虑到同一地理区域内不同方言语种间的相似性及不同区域方言语种的相关性,首次基于 MFCC 特征,以方言区域为辅助任务,构建基于参数硬共享方言语种识别模型;

(2)发现基于参数硬共享的模型仅在任务相关性强时具有良好性能,而方言分类任务的相关性不能被明确界定。由此,首次提出基于参数稀疏共享的多任务方言分类模型,该模型通过联合训练,自主确定任务间的相关性,并对子任务部分网络进行自适应共享。

1 相关研究

1.1 语音语种识别

语音语种识别是模式识别的一个重要分支。其首要任务是从输入声音中提取特征作为多维向量。特征提取是通过将语音波形以相对最小的数据速率转换为参数表示形式进行后续处理和分析来实现。梅尔频率倒谱系数(MFCC)[3]、线性预测倒谱系数(linear prediction cepstral coefficients, LPCC)[4]、离散小波变换(discrete wavelet transformation, DWT)[5]和感知线性预测(perceptual linear predictive, PLP)[6]是常见的语音特征参数。这些方法已经在广泛的应用中进行了测试,具有很高的可靠性和可接受性。方言研究通常将语音预处理为上述特征参数。

语音语种识别其主流起先是人工神经网络,继而高斯混合模型、HMM、MVC 等模型也被广泛应用。随后发展到基于神经网络的应用。2011 年 Ge 等人提出基于情景依赖的预训练 DNN 方法[7],蒋兵在 2015 年提出基于深层神经网络提取音素相关深瓶颈特征(deep bottleneck feature, DBF)的语种识别方法[8]。CNN 在语音识别领域有广泛的应用,Koller 等提出了混合 CNN-HMM 模型,关注到语音输入的连续性[9]。2017 年,微软在新的对话语音识别系统中加入了 CNN-BLSTM 声学模型。这些均取得了不错的成果。近几年,循环神经网络(recurrent neural network, RNN)也广泛应用到了语音识别领域[10]。LSTM 和 GRU 网络也被相继提出,由于其改善了 RNN 所存在的长期依赖问题,对序列化的语音特征的识别有良好性能。如使用 CNN 与 GRU 融合深度神经网络对语音语种进行分类。相关研究还在此基础上对原有 LSTM 及 GRU 模型进行改进,以达到性能优化的目的[11]。

近年来,方言识别多采用端到端的识别模型,通常使用卷积层和循环层结合的网络[12]。文献[2]利用卷积神经网络对江苏省县级市方言进行分类识别,文献[13]将多任务学习运用到解决方言语种识别的问题上,但其使用语音音频时段较长。

1.2 基于多任务学习的语音识别研究

多任务学习(multitask learning, MTL)假设不同任务数据分布之间存在一定的相似性,在此基础上通过共同训练和优化建立任务之间的联系。充分促进任务之间的信息交换并达到了相互学习的目标。各个子任务可以从其他任务获得一定的启发,并间接利用其他任务的数据,达到了提升各自任务学习性能的目的[14]。

S Ruder 提出了深度学习中 MTL 的两种最常用的方法。阐明了 MTL 原理,为自然语言处理和语音识别提供指南。

近年来,多任务学习及 LSTM 循环网络的提出为方言语种识别提供了研究的空间。该文应用 LSTM 网络构建单任务学习模型,对其参数进行优化。进一步利用不同地域方言的相关性,提出基于参数软共享的多任务模型,共享隐藏层信息提高准确率。考虑到同一方言区域的方言语种的相似性,以方言区域为辅助任务提出基于参数硬共享的多任务学习模型,提升模型的泛化性。继而为确定任务相关性,提出基于参数稀疏共享机制的多任务学习模型。在方言语种的选取上,很多研究多选取印地方言[15],某一单一区域方

言如湖南方言等[16],部分语音识别相关研究选取少数民族语言如维吾尔语、藏语等。该文选取涵盖 6 大方言区域的 10 种中国方言语种。

2　语音特征

音频中包含丰富的语音特征信息,不同的特征向量往往可以表征不同的声学意义,由此抽取语音特征,挖掘语音特征关联信息,对语种识别至关重要。语音特征提取是指从一段语音中选择有效的音频表征。原始语音信号是典型的非平稳信号,而语音特征提取的分析手段是针对平稳信号的。由此通常假定 10 ms ~ 30 ms 的短时内信号为平稳信号,并基于此进行成分分析。该文对原始音频的长语音信号进行加窗、分帧处理,进而获得多个平稳的短时信号。在此基础上,进行语音特征的分析提取。该文抽取 MFCC 特征。

图 1 展示了 MFCC 特征的抽取方法及其与常见语音特征 Fbank、语谱图的对比。

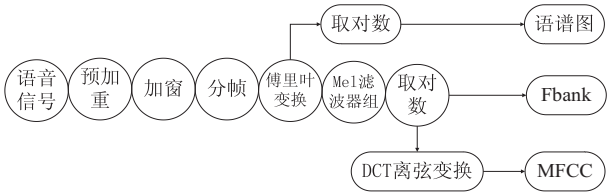


图 1　MFCC 抽取方法及其与 Fbank、语谱图的对比

语谱图通过将语音信号进行傅里叶变换后取对数而得,保存了更多的原始语音信号信息但噪音较多。Fbank 特征是在对语音信号进行傅里叶变换后,进一步使用梅尔滤波器组进行处理,并对结果取对数而得。MFCC 和 Fbank 参考人类对声音高低的敏感度,利用梅尔刻度对这种敏感度进行量化表示。其获取流程与人耳处理声音的流程类似。MFCC 特征是在 Fbank 基础上进行 DCT 离弦变换后获得的倒谱参数,因此数据维度小。

3　基于单任务学习的方言分类模型

本节基于 MFCC 特征构建单任务神经网络模型。模型构建包含数据采集、特征提取及模型训练三个过程,训练过程如算法 1 所示。训练及测试数据集来自于科大讯飞 AI 挑战大赛所提供的方言数据集。

3.1　特征提取

通过 Python wave 库提取语音信号 MFCC(mel-frequency cepstral cofficients),并将 10 种方言数据处理成[data,label](数据及其对应标签,其中闽南、客家、上海、合肥、陕西、宁夏、长沙、河北、南昌、四川依次对应标签 0 ~ 9)。

3.2　基于单任务神经网络的方言语种分类

该模型由 CNN 和 LSTM 两部分组成,网络结构如表 1 所示,其中 C1、C2 为卷积层,D1、D2 为全连接层。以 C1 参数为例,(3×3,1)中,3×3 表示卷积核大小,1 表示步长,8 表示卷积核个数。上述三种单任务神经网络模型实验结果及性能分析见 5.3 节。模型训练过程见算法 1。

表 1　基于 MFCC 的单任务模型架构

层	描述	参数
Input	——	13×199
C1_Conv	卷积层	(3×3,1),8
C2_Conv	卷积层	(3×3,1),16
LSTM	LSTM 层	40
D1_Dense	全连接层	40,128

算法 1:单任务语种分类模型训练算法。

输入:训练集数据及标签(Train_Data,Train_Label)、测试集数据及标签(Dev_Data,Dev Label)、训练轮数(Epoch)

```
#初始化
Epoch=N;
Best_acc=0;
For i=1 to N
#将数据输入网络进行训练
Train_Pre= Net(Train_Data);
#对网络输出执行 softmax 运算
Train_Pre=Softmax(Train_Pre);
#计算训练集准确率及 Loss 函数值
Train_acc=Acc(Train_Pre, Train_Label)
Loss=Loss (Train_Data, Train_Label)
#反向传播更新权重
Loss.backward()
Dev_Pre=Net(Data);
Dev_Pre=Softmax(Dev_Pre);
Dev_acc=Acc(Dev_Pre, Dev_Label)
Best_acc=Dev_acc
输出(Train_acc,Dev_acc)
IF Dev_acc> Best_acc:
Best_acc=Dev_acc
保存模型
EDN IF
输出模型;
```

输出:训练集准确率(Train_acc) 测试集准确率(Dev_acc)

4　基于多任务学习的方言分类模型

多任务学习(Multi-task learning,MTL)在自然语言处理和计算机图像[17]等领域取得了一定的成功。其核心是通过同时训练多个相关任务并共享不同任务间的参数以提升模型整体的泛化能力,多任务学习包含参数软共享和参数硬共享两种参数共享方式。本节基于 MFCC 特征构建基于参数软共享的多语种任务方言分类模型以及基于参数硬共享的以方言区域分类

为辅助任务的方言分类模型。

4.1 以方言区域识别为辅助任务的方言分类模型

本节介绍以方言语种分类为主任务、以方言区域识别为辅助任务构建的基于参数硬共享的多任务学习模型。中国方言可划分为十大方言,分别是:官话方言、客家方言、晋方言、徽方言、湘方言、吴方言、粤方言、闽方言、赣方言、平话土话。不同方言区域有其各自特点,与此同时不同方言区域特征也有一定的交叉性,如赣方言长期受地理位置邻近的客家方言的影响,导致其特征与客家方言相似。本节以方言区域分类做辅助任务,为主任务提供不同方言区域特征信息和方言区域的交叉信息,由此构建基于参数硬共享的多任务学习模型。参数硬共享机制共享任务间的隐藏层,并保留各自输出层,适合任务相关性较高的情况。以方言区域识别为辅助任务的方言语种分类模型,主任务方言语种分类和辅助任务区域识别任务共享同一网络架构,共享各个方言区域的隐藏信息,保留主任务和辅助任务的输出,网络联合训练所有任务损失值,通过输出评估结果。方言区域与标签对应关系如表 2 所示,实验结果见 5.4.1 节。

该模型网络结构同单任务模型相同,模型结构如图 2 所示。该模型通过在共享层实现参数硬共享,计算主任务和辅助任务的损失 $loss_1$, $loss_2$ 并进行加权求和,权重均为 0.5,如公式(5)。对 $loss_{sum}$ 进行反向传播和更新,训练并保存模型。

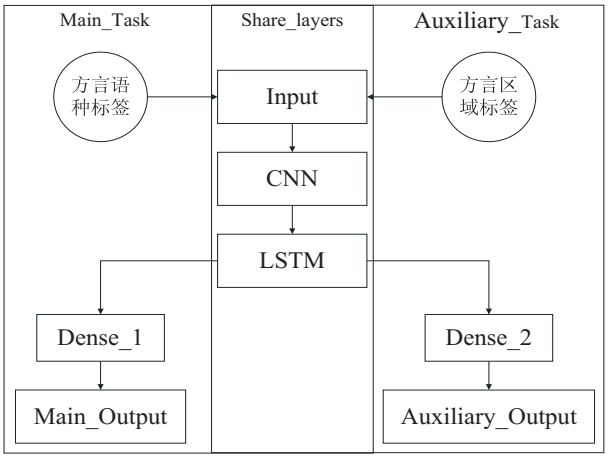


图 2 基于参数硬共享的模型架构图

表 2 方言区域标签对应表

方言	方言区域	标签
闽南	闽方言	0
客家	客家方言	1
上海,合肥	吴方言	2
陕西,宁夏,河北,四川	官话方言	3
南昌	赣方言	4
长沙	湘方言	5

4.2 基于稀疏参数共享的方言分类模型

本节介绍基于稀疏参数共享[18]的方言分类模型,稀疏共享通过自主发现子任务间的共性,裁剪不重要的网络参数,完成子任务间部分参数的共享,在提高共享机制灵活性的同时减少模型参数量,提升模型执行效率。

训练流程包含 3 步,首先,构建一个超参数模型,称为基础网络(base network)。然后,子任务基于 base network 开始训练,训练过程中子任务根据需求动态裁剪冗余参数。最后,从基础网络中抽取出各个任务需要的子网络。由于是自动训练,往往任务相关性高的子网络会共享部分网络权重,而任务相关性不高的子网络独自为一体。与硬共享不同,稀疏共享因为子任务间部分网络的重叠,实现了参数共享的目的。图 2 所示为所用的稀疏共享机制。首先,选用第四节构建的单任务模型作为稀疏共享机制的基础网络,并定义方言语种分类和方言区域分类两个子任务,其中 Task1 为方言语种分类任务、Task2 为方言区域分类任务。进一步,使用二进制掩码矩阵 $\boldsymbol{M} \in \{0,1\}$ 为各个任务选择子网,图中 Mask1、Mask2 为各个任务子网络,将各个任务子网络合并在一起构成了稀疏共享结构。然后,分别使用两个子任务数据进行联合训练,联合训练时抽取各个任务的数据对相应子网络进行训练,基础网络中子任务重合的参数被多次训练。参数的稀疏共享机制既能够体现出两种分类任务之间的差异性,又注重语种与其所属语言区域特征的交叉性与相关性。

4.2.1 子任务网络

子任务网络通过对 base network 进行迭代幅度裁剪(iterative magnitude pruning, IMP)[19]而得,训练过程如算法 2 所示。网络结构和参数由二进制掩码 Mask 矩阵和基础网络决定。设基础网络的参数为 θ,子任务 t 的 Mask 矩阵为 $\boldsymbol{M}_t$,则子任务 t 的网参数为 $\boldsymbol{M}_t \cdot \theta$。$\alpha$ 为每轮训练过程中子网参数权重保留率(percent of weights remaining),即子网中 α 的参数的权重被保留,将权重由小到大排列,前 $1-\alpha$ 的参数权重被裁剪,若权重被保留则其对应的 $\boldsymbol{M}_t$ 值为 1,反之为 0。

算法 2:稀疏共享模型子网络训练算法。

输入:基本网络,最低保留率 Min α;

输出:最高准确率对应的子网络 $\boldsymbol{M}_t$。

模型参数初始化

For $t = 1 \cdots t$

Do

初始化 $\boldsymbol{M}_t$

初始化保留率 $\alpha = 100\%$

初始化网络参数总数 total_params

Do
Train( )
Dev( )
计算保留参数个数 total_m
#更新保留率
α =1-round (100% * total_m/self. total_params,2)
#裁剪并保存网络 θ ,mask,acc
While α >Min(α) or pruning epoch>10

4.2.2　并行训练子任务网络

每个任务训练时都只用到了对应的子网络,但各个子任务的网络会共享网络参数,因此这部分参数会被多个任务的数据进行更新,训练过程如算法 3 所示。实验结果详见 5.4.2 节。

算法 3:并行训练子网的算法。
输入:基本网络,子任务数据集,子网络 Mask;
输出:测试集准确率。
Epoch= N ;
For i =1 to N
选择任务 t
$\boldsymbol{M} = \boldsymbol{M}_t$
#为任务 t 随机选择一个 Batch 的数据
Train( )
Dev( ) #更新子网络参数
Acc( ) #计算相应任务准确率

5　实验分析

5.1　实验平台

硬件平台:使用高性能服务器进行模型训练,该平台配置了 Intel(R) Core(TM) CPU,64 GB 内存,2 块 GeForce RTX 2080 Ti GPU。

软件平台:服务器运行 Ubuntu 系统,搭载 Pytorch(1.6.1)、CUDNN(1.6.0)、CUDA(10.0)。模型均基于 Pytorch 框架进行实现。

5.2　数据集

实验数据集为科大讯飞 AI 开发者大赛提供的方言数据。数据集共包括十种方言,每种方言包含 40 个本地人的朗读风格语音数据,数据以采样率 16 000 Hz,16 比特量化的 PCM 格式存储。数据集包含训练集和测试集两个部分。训练集每种方言有 5 000 句语音,包含 15 位男性和 15 位女性,共 30 位说话人,每个说话人 200 句语音;测试集每种方言包含 5 个说话人,其中 3 名女性和 2 名男性。训练集、测试集的说话人均没有重复,训练集方言音频条数为 6 000 条,测试集方言音频条数为 500 条,同时,对所使用的数据集进行了统一时长及去噪处理。

统一时长:由于数据音频时长不统一,且部分音频中存在大量的空白时段,通过音频裁剪或填充方法,将所有音频数据处理为 2 秒。

去噪:实验使用谱减法对数据进行去噪。

5.3　单任务方言分类模型结果分析

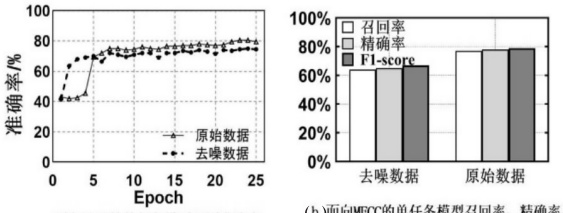
分别抽取数据集的 MFCC 的语音特征,并分别输入单任务模型进行训练。代价函数为交叉熵损失函数,epoch 为 25。

采用少量数据对单任务模型进行参数调整。参数调整结果见表 3。

表 3　参数调整对准确率影响变化趋势

参数	调整策略		准确率
LSTM 层数	1	2	↑
隐藏节点个数	128	256	↑
归一化	√	×	↓↓
剔除有效信息含量较低的数据	保留空白占比小于 20% 的数据	保留空白占比小于 30% 的数据	↓

在表 3 准确率变化栏中,准确率的变化趋势由符号‘=’,‘↑’,‘↓’表示,符号数量表示程度,‘=’表示调整参数后准确率变化不大,‘↑↑’表示调整参数后准确率提高幅度大,‘↓↓’反之。分别将原始数据和去噪数据输入优化后的网络模型,记录测试集准确率。如图 3(a)所示,使用原始数据训练的模型准确率可达 79.04%,去噪数据准确率达到 76%。其召回率、精确率、F1_score 值如图 3(b)所示。



(a)面向MFCC的单任务模型识别准确率　(b)面向MFCC的单任务模型召回率、精确率、F1_score

图 3　面向 MFCC 的单任务模型识别准确率及召回率、精确率、F1_score

综上可知,三种单任务模型在去噪数据集上的准确率低于原始数据训练模型。这是由于降噪过程中需要对原始音频数据进行平滑处理,该处理方式有利于改善用户听感,但另一方面,平滑处理使频谱信息变得模糊,导致部分原始音频信息丢失,从而无法获得较高准确率。后续实验模型均基于原始数据进行构建。

5.4　多任务语种识别模型结果分析

5.4.1　以方言区域识别为辅助任务的硬共享方言分类模型

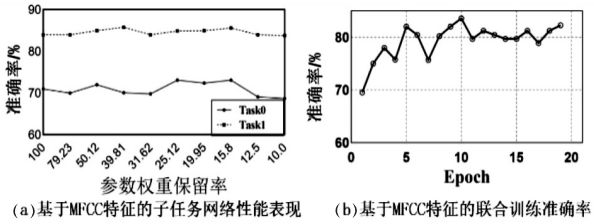
首先,为数据集中每条音频信息进行区域标记。形式为[data,label1,label2],其中 label1 表示数据方言语种标签,label2 表示数据的方言区域标签,然后,将 MFCC 特征原始数据输入模型(模型结构见表 2),实验结果如图 3 所示。相比于单任务模型,主任务和辅助任务联合训练的参数硬共享模型准确率平均提升

1%。其在测试集上的准确率达 79.9%。模型召回率为 79.1%,精确率为 77.4%,F1_score 为 76.7%。

5.4.2 基于稀疏共享的方言分类模型

本节选择单任务模型作为 base network。分别选取最优的子任务模型,Task0 为方言语种分类任务,Task1 为方言区域分类任务。图 4(a)为不同参数权重保留率下子任务的准确率,实验选择准确率最高的子任务模型。子任务模型选择情况为任务 Task0 选取参数保留率为 25.12% 的模型,Task1 选取参数保留率为 39.81% 的模型,

对不同任务分别进行联合训练,其准确率如图 4(b)所示。召回率为 77%,精确率为 76%,F1_score 为 65%。模型准确率可达 83.59%,较基于参数硬共享的方言识别模型准确率提高 3%。



(a)基于MFCC特征的子任务网络性能表现 (b)基于MFCC特征的联合训练准确率

图 4 基于 MFCC 特征的子任务网络性能表现及联合训练准确率

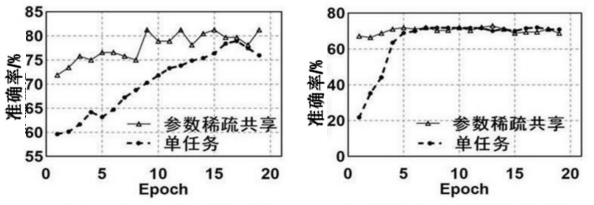
5.4.3 基于不同语音特征的多任务模型结果分析

为进一步验证稀疏共享分类模型的有效性和优越性,在原始数据集上提取语谱图,Fbank 两个常见的语音特征,并针对两种特征建立基于 LSTM 的单任务模型,进一步构建基于参数硬共享和基于参数系数共享的多任务模型,其准确率如图 5(a)、5(b)所示。语谱图、Fbank 特征在参数稀疏共享模型上的准确率分别提升 3%、1%,召回率、精确率、F1-score 如表 4 所示。

表 4 模型召回率、精确率、F1_score 对比表

特征	模型	召回率	精确率	F1-score
语谱图	单任务	73	75	71
	参数稀疏共享	75	76	71
Fbank	单任务	70	71	69
	参数稀疏共享	69	72	70

语谱图与 Fbank 特征在多任务模型上的表现均优于单任务模型。基于在稀疏共享的多任务的模型上获取最高准确率。



(a)基于语谱图的模型准确率对比 (b)基于Fbank的模型准确率对比

图 5 基于语谱图、Fbank 的模型准确率对比

5.5 实验对比

将基于稀疏共享的方言分类模型(sparse-sharing network,SSNet)同文献[14]提出的 ATLNet 方言分类模型进行比较。为了保证数据的统一性,用同样的 2 秒数据集训练 ATLNet 网络。由于 ATLNet(准确率仅为 78.3%)模型无法充分挖掘方言任务间相关性,由此在 2 秒数据集中,提出的基于稀疏共享的方言分类模型性能更优,准确率达 83.59%。

5.6 实验结果总结

针对中国方言的复杂性、相似性的特征,结合语音数据特征特性,首先应用 LSTM 搭建面向 MFCC 的方言语种识别系统,并对模型进行调参优化。进一步提出基于参数硬共享、稀疏共享的两种多任务模型。并以 Fbank、语谱图为输入进行有效性验证,获得表现最优(准确率达 83.59%)的基于稀疏参数共享的方言分类模型。

6 结束语

中国地域辽阔,方言种类繁多,汉语方言识别一直是语音识别中的难点,提升方言分类准确率对于下一步的方言识别至关重要。首先抽取典型的语音特征 MFCC 进行对比研究,并基于该语音特征分别构建 LSTM 方言分类模型,准确率可达 79.04%。进一步挖掘中国十种方言的地域特征,建立以方言区域为辅助任务的参数硬共享多任务方言分类模型,由于多个子任务之间存在一定相关性,相比单任务模型,联合训练多任务模型可以挖掘到更多的数据特性,由此也获得了更好的性能,准确率可达 79.9%。最后,为了进一步提升多任务模型规模,删除冗余的网络节点对结果的影响,提出基于参数稀疏共享的多任务方言分类模型,联合训练多个子任务的过程中,对 base network 进行自动裁剪,并共享相关度高的网络权重,在降低网络复杂度的同时,提升识别性能,准确率可达 83.59%。

综上,MFCC 语音特征,利用三种不同的网络结构对方言分类进行建模研究,其中基于参数稀疏共享多任务方言分类模型性能最优。下一步,作者将从以下三个方面进行进一步研究:(1)针对多任务模型中不同子任务 loss 值对模型性能的影响进行研究;(2)研究 Attenion 机制对方言分类的影响;(3)利用迁移学习降低模型训练开销。

参考文献:

[1] OTTER D W,MEDINA J R,KALITA J K. A survey of the usages of deep learning in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,32(2):604-624.

[2] 魏黎明. 基于卷积神经网络的方言分类方法研究[D]. 南京:东南大学,2018.

[3] KODUKULA S R M, YEGNANARAYANA B. Combining evidence from residual phase and MFCC features for speaker recognition[J]. IEEE Signal Processing Letters, 2005, 13(1):52-55.

[4] YUAN Y, ZHAO P, ZHOU Q. Research of speaker recognition based on combination of LPCC and MFCC[C]//2010 IEEE international conference on intelligent computing and intelligent systems. Berlin, Germany: IEEE, 2013: 765-767.

[5] SHENSA M J. The discrete wavelet transform: wedding the a trous and Mallat algorithms[J]. IEEE Transactions on Signal Processing, 1992, 40(10): 2464-2482.

[6] HERMANSKY H. Perceptual linear predictive (PLP) analysis of speech[J]. The Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752.

[7] DAHL G E, YU Dong, DENG Li, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(1): 30-42.

[8] 蒋　兵. 语种识别深度学习方法研究[D]. 合肥:中国科学技术大学,2015.

[9] KOLLER O, ZARGARAN S, NEY H, et al. Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs[J]. International Journal of Computer Vision, 2018, 126(2): 1311-1325.

[10] MIAO Y, GOWAYYED M, METZE F. EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding[C]//2015 IEEE workshop on automatic speech recognition and understanding (ASRU). Piscataway, NJ: IEEE, 2015: 167-174.

[11] SHEWALKAR A, NYAVANANDI D, LUDWIG S A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU[J]. Journal of Artificial Intelligence and Soft Computing Research, 2019, 9(4): 235-245.

[12] ALI A R. Multi-dialect Arabic speech recognition[C]//2020 international joint conference on neural networks (IJCNN). Piscataway, NJ: IEEE, 2020: 1-7.

[13] 秦晨光,王　海,任　杰,等. 基于多任务学习的方言语种识别[J]. 计算机研究与发展,2019,56(12):2632-2640.

[14] RUDER S. An overview of multi-task learning in deep neural networks[J]. arXiv: 1706.05098, 2017.

[15] STEVENS K N. Toward a model of lexical access based on acoustic landmarks and distinctive features[J]. The Journal of the Acoustical Society of America, 2002, 111(4): 1872-1891.

[16] KIM S, HORI T, WATANABE S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]//2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 4835-4839.

[17] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//2018 IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2018: 7482-7491.

[18] SUN T, SHAO Y, LI X, et al. Learning sparse sharing architectures for multiple tasks[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto, CA: AAAI, 2019: 8936-8943.

[19] ELESEDY B, KANADE V, TEH Y W. Lottery tickets in linear models: an analysis of iterative magnitude pruning[J]. arXiv: 2007.08243v2, 2020.