文章编号:1671-6833(2023)04-0001-09

专栏:机器学习与计算应用

【主持人】张震:大数据协同安全技术国家工程实验室副主任

【按 语】机器学习是人工智能的一个重要分支,涉及概率论、统计学、算法复杂度理论等多门学科。通过让计算机从大量数据中自动学习和提取规律,使其具备类似于人类的学习和决策能力。近年来,物联网、云计算和大数据等技术的广泛应用为机器学习的研究提供了强大的支持。《国务院关于印发新一代人工智能发展规划的通知》指出:人工智能成为国际竞争的新焦点、经济发展的新引擎。教育部发布的《高等学校人工智能创新行动计划》倡导:面对新一代人工智能发展的机遇,高校要进一步加强应用基础研究和共性关键技术突破,要不断推动人工智能与实体经济深度融合、为经济发展培育新动能。科技部专门加强人工智能顶层设计,启动实施新一代人工智能重大科技项目。在此背景下,对于机器学习与计算应用的研究具有重要的理论意义和实际应用价值。在论文《低资源少样本连续语音识别最新进展》中,作者系统地总结分析了低资源少样本语音识别技术的最新技术、研究难点和未来的研究方向;在论文《基于对抗机制的彩色图像隐写分析算法》中,作者提出了一种基于对抗机制的深度卷积网络,有效解决了彩色图形的隐写分析问题;在论文《改进 YOLOv5 算法在停车场火灾检测中的应用》中,作者针对停车场火灾检测的应用场景,提出了一种改进的 YOLOv5 算法,可有效检测出小型火焰目标;在论文《复合可靠性分析下的不平衡数据证据分类》中,作者针对传统分类模型在处理不平衡数据时会侧重于大类而忽略小类的问题,提出了一种复合可靠性分析下的不平衡数据证据分类方法,提升了模型对不平衡数据的分类能力。

低资源少样本连续语音识别最新进展

屈 丹, 杨绪魁, 闫红刚, 陈雅淇, 牛 铜

(战略支援部队信息工程大学 信息系统工程学院,河南 郑州 450001)

摘 要: 低资源少样本语音识别是目前语音识别行业面临的迫切技术需求。首先,总结了低资源连续语音识别技术的框架技术,重点介绍了低资源语音在特征提取、声学建模和资源扩展等方面的若干关键技术研究进展。其次,在连续语音识别框架技术发展的基础上,重点阐述了生成对抗网络、自监督表示学习、深度强化学习和元学习等高级深度学习技术在解决少样本语音识别方面的最新发展,如 FGSM、wav2vec、AMS 等代表性方法。在此基础上,分析了目前该技术面临的互补有限、数据和任务不均衡与模型轻量化部署问题。最后,对低资源少样本连续语音识别进行了总结,提出未来少样本训练识别的研究方向可以朝着先验信息引入、假设空间约束条件设定等方向进一步研究。

关键词: 低资源少样本; 连续语音识别; 生成对抗网络; 自监督表示学习; 深度强化学习; 元学习

中图分类号: TN912.34 **文献标志码:** A **doi:** 10.13705/j.issn.1671-6833.2023.04.014

收稿日期: 2023-03-20;**修订日期:** 2023-04-05

基金项目: 国家自然科学基金资助项目(62171470);河南省中原科技创新领军人才项目(234200510019)

作者简介: 屈丹(1974—),女,吉林九台人,战略支援部队信息工程大学教授,博士,博士生导师,主要从事人工智能理论与智能信息处理研究,E-mail:qudanqudan@sina.com。

引用本文: 屈丹,杨绪魁,闫红刚,等. 低资源少样本连续语音识别最新进展[J]. 郑州大学学报(工学版),2023,44(4):1-9.(QU D, YANG X K, YAN H G, et al. Overview of recent progress in low-resource few-shot continuous speech recognition[J]. Journal of Zhengzhou University (Engineering Science),2023,44(4):1-9.)

自动语音识别技术(automatic speech recognition, ASR)是人工智能领域非常活跃的一个领域。传统的语音识别框架通过DNN或GMM与HMM联合实现混合结构的声学模型,并需要语言模型和发音词典的配合才能完成语音识别任务,该框架的缺点在于模型结构复杂、训练步骤烦琐和专家知识依赖性强。为了解决这些问题,研究人员逐渐开始研究下一代语音识别框架——端到端(end-to-end)的语音识别技术。

2006年,Graves等[1]首次提出端到端连接时序分类(connectionist temporal classification, CTC)算法,并用其训练了一个由RNN编码器和线性分类器构成的深度网络,实现了端到端语音识别。2013年,Graves等[2]又提出了RNN转换器(RNN transducer, RNN-T)语音识别模型。Chan等[3]提出了"listen, attend, and spell (LAS)"模型,LAS模型的listen、attend和spell分别表示编码器、注意力机制和解码器,通过注意力机制在编码器输出的特征序列上构造上下文相关矢量,并与前一时刻的解码器的输出一并用于生成当前时刻的输出,该模型超过之前的模型,达到了当时最好的性能。2017年,谷歌公司在文本处理领域首先提出了基于全注意力的变换器(Transformer)模型[4],研究者将其扩展到语音识别领域,在训练效率和识别性能上达到了新的最佳性能,逐渐成为端到端语音识别领域的基本模型。随后很多针对Transformer的改进模型陆续被提出[5]。

当代最新语音识别技术在某些特定数据集上已突破了人类的能力极限,超过了人类听抄的最高水平。但这种突破基于两个特定条件:一是信道环境较为理想,即通常话音质量非常好、采集场地相对单一、噪声起伏变化不大;二是训练语料非常充足,尤其是对于汉语、英语等大样本语言,其工业级实用系统训练所需的标注数据常常超过数千小时甚至是数万小时。但在大多数特定行业应用领域,所面临的实际条件对少样本机器学习提出了更为迫切的需求,主要表现在以下两个方面。一是现实环境极其复杂且难以预测,信道条件往往多种多样,噪声也起伏变化差异较大,如既有录播室环境的高保真信号,也有经过电信网络或者其他带噪通道传输的窄带信号,这种复杂的条件为语音处理与识别处理带来诸多难题。二是标注数据匮乏使得传统机器学习难以形成有效认知。无论是哪种语言的语音识别系统都需要有大量标注数据。世界上有超过7 000种语言,而其中仅有为数很少的几种语言(如汉语普通话、英语等)具有充足的标注数据,除此之外的绝大部分语言或方言都面临"资源匮乏"或"少样本"困境。而往往这些"少样本"语言如马来语、印尼语、印地语、土耳其语等在国家战略交流中也非常重要。很多行业应用的现实环境就不具备大规模采集标注数据的条件,使得语音识别行业应用面临的迫切需求是突破少样本机器学习的性能极限、探讨更先进的技术方法、从理论和技术上攻克难题。

1 低资源连续语音识别框架技术

近年来,国内外针对低资源少样本下的语音识别问题进行了大量研究。其中以2011年美国情报先进研究计划署(Intelligence Advanced Research Projects Activity, IARPA)的Babel计划[6],荷兰代尔夫特理工大学、爱尔兰都柏林城市大学等组织发起的每年一次的多媒体评测(MediaEval)为代表。低资源小语种语音识别指缺乏用于训练的相关数据资源,包括标注语音、发音字典和文本等,其中尤其以发音字典和标注语音影响最大。小语种语音数据难以获取不仅表现在语音上,更表现在语料、发音词典和标注资源上。与传统语音识别相比,低资源少样本连续语音识别有许多针对性的技术,主要体现在如下3个方面,总体框架如图1所示。

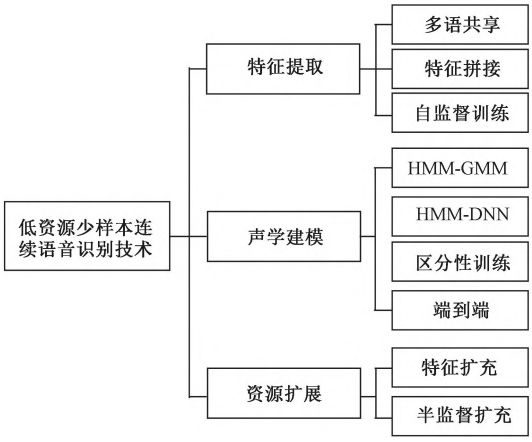

图1 低资源少样本连续语音识别技术

Figure 1 Low-resource few-shot continuous speech recognition

1.1 低资源少样本特征提取技术

深层特征具有较强的鲁棒性,在环境、说话人发生变动的情况下不确定性更小,因此目前广泛采用深度神经网络(deep neural network, DNN)提取更加稳健的高层语义表示特征[7]。2011年前后,在深度学习引入语音识别领域不久,许多研究关注于利用DNN提取深层特征,典型的研究成果有tandem特征和瓶颈(bottleneck, BN)特征。目前关于低资源条件下的特征提取技术研究主要集中在3个方面:

一是利用多语数据通过共享神经网络权值的方法实现相关特征的提取,从而提高在其基础上建立的语音识别系统的性能;二是采用不同声学特征参数拼接融合的方法增强特征的区分性和稳健性[8];三是借助一些无监督或者自监督方法训练编码器,可以将编码器的输出作为特征映射函数,将信号或特征经过编码器后的输出作为特征,或者将编码器借助特定领域数据进行迁移后作为特征提取模型来完成特征映射,该部分内容将在 2.2 节的自监督表示学习部分中重点阐述。

1.2　低资源声学建模技术

语音识别中的声学建模技术发展可以分成 4 个阶段:一是传统隐马尔可夫模型-高斯混合模型(hidden Markov model Gaussian mixture model, HMM-GMM);二是 DNN 混合声学模型建模技术;三是目标函数的区分性训练准则;四是打破传统框架的端到端建模技术。

在 HMM-GMM 框架模型方面,2011 年 Povey 等[9]提出子空间高斯混合模型(subspace Gaussian mixture model, SGMM)来对 HMM 发射概率进行建模;Imseng 等[10]进一步利用 Kullback-Leibler 距离对传统 GMM 进行正则化。在 DNN 混合声学模型建模技术方面,Mohamed 等[11]提出使用 DNN 代替 HMM-GMM 声学模型中的 GMM,用于对发射概率进行建模,大大减少了对标注数据的依赖性。后续的工作对 DNN 的结构进行了进一步的探索,典型的如时间延迟神经网络(factorized time-delay neural network, TDNN-F)[12-13]等。此外,为了进一步提升模型表现力,深度学习的区分性训练准则如最大互信息(maximal mutual information, MMI)、提升最大互信息(boosted maximal mutual information, BMMI)、最小音素/词错误(minimum phone/word error, MPE/MWE)、状态级最小贝叶斯风险(state-level minimum Bayes risk, sMBR)、词格无关最大互信息(lattice-free maximum mutual information, LF-MMI)[14]等,也被引入到模型训练中,以进一步提升识别率。鉴于混合模型需要多个模型分别训练,优化难度大,且对专业知识需求高,研究者越来越关注端到端的结构,目前已成为主流的学习范式。但是这也并不意味着丢弃之前的研究,类似于区分性训练准则也被广泛整合到端到端模型的训练过程中。

1.3　低资源语音识别资源扩展技术

标注数据匮乏是低资源的一个显著特点,因此需要研究者考虑扩展数据集。目前由于语音标注资源比较受限于高成本,因此主要考虑标注语音样本扩展技术。训练数据扩展有两种策略。一种是不改变文本标注,只对音频或声学特征进行扩展。最简单的做法是针对现有训练数据,在保证基本语义不变的情况下通过改变语速等方法获得额外的训练数据。因此噪声添加、声道长度扰动技术[15]、语速扰动(speed perturbation, SP)方法[16]作为经典数据拓展方法常常被用于低资源语音识别中。后期更多的数据增强策略被采用,包括 SpecAugment[17]、Wav-Aug[18]、MixSpeech 等。目前这些方法已经成为语音识别模型中的默认配置,在各种场景、语种下展现出较强的鲁棒性,可以使识别效果得到持续的提升。

另外一种重要的方法是基于半监督的数据扩充。首先,利用训练数据训练一个初始的语音识别模型。其次,对无标注的单语数据进行识别,将识别结果作为这些单语数据的标注,一般称之为伪标注数据。最后,根据一定的策略对这些伪标注数据进行筛选,将筛选后的伪标注数据和原始训练数据混合,重新训练整个模型。比较典型的有噪声学生训练(noisy student training, NST)[19]。但是该方法一方面需要多轮重新训练,会耗费大量训练资源;另一方面其性能依赖于好的标签选择策略。

2　少样本语音识别中的高级深度学习技术

为了能够在受限标注样本情况下获得更好的性能,近年来深度学习中的高级建模技术也为低资源语音识别技术带来了新的活力,主要技术包括生成对抗网络、深度强化学习、迁移学习、元学习、自监督表示学习等,其核心目标在于从几个方面克服低资源恶劣环境的影响:一是数据层面的拓展,例如生成对抗网络和深度强化学习都可以进行数据增强;二是利用已有数据获得更有泛化力和表征性更强的特征提取算法,如自编码和自监督表示学习技术;三是已有相关知识的借鉴和利用,如迁移学习和元学习技术;四是寻找一些更好的学习机制,例如元学习、对抗学习和深度强化学习机制等。总体框架如图 2 所示。

2.1　生成对抗网络

生成对抗网络(generative adversarial networks, GAN)是蒙特利尔大学 Goodfellow 等[20]在 2014 年提出的一种生成式模型。GAN 的基本思想来源于博弈论中的二人零和博弈[21],通过从训练库中获取多个真实训练样本,利用博弈对抗思想学习这些样本的生成概率分布。GAN 模型中的两个博弈方分别由生成器 G 和判别器 D 构成。其优化目标为

$$\min_G \max_D V(D,G) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + E_{z\sim p_z(z)}[\log(1 - D(G(\boldsymbol{z})))]。\quad (1)$$

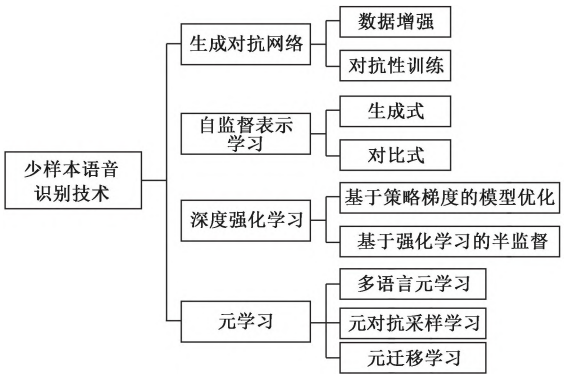

图2 少样本语音识别中的高级深度学习技术框架

Figure 2 Advanced deep learning framework in few-shot speech recognition

生成器在噪声先验分布 p_z 中采样生成样本以捕捉样本数据的分布,判别器用于最大化式(1)的训练样本来自于训练数据 p_{data}(而非生成数据)的概率,二者都可以采用结构多样的模型来完成。生成对抗网络在低资源语音识别中的应用主要体现在两个方面:一是利用对抗网络来产生训练数据等实现数据增强和拓展;二是利用对抗网络进行对抗性训练以减少训练环境和测试环境失配。

在数据增强方面,2019年,Qian等[22]利用GAN进行数据生成以改善噪声条件下的语音识别,一方面利用生成对抗网络构建无监督学习框架进行声学建模,另外一方面利用条件GAN生成带标签信息的真实语音用于声学模型建模。实验证明了这种数据增强可以降低噪声环境下语音识别系统的词错误率。2018年,Sun等[23]采用快速梯度符号法(fast gradient sign method, FGSM)生成对抗样本进行数据增强,与一些静态转换数据增强方法不同,样本可以基于当前声学模型参数动态产生。该方法在噪声和变化信道条件下获得了较好的性能。

在对抗性训练方面,2016年,Shinohara[24]提出对抗多任务训练(adversarial multi-task learning, AMT)方法。该方法最早将GAN用于语音识别,其基本思想是基于DNN高层特征对噪声具有鲁棒性的特点,将语音经DNN编码器处理后的高层次特征送入到GAN网络中进行对抗性训练来实现对音子建模。生成器目标在于最大化分类概率,判别器目标在于最大化信号域分类概率,而编码器目标在于上述两种目标和。实验结果表明该方法可提升噪声条件下语音识别系统性能。2018年,Liu等[25]提出了一种直接增强声学模型噪声鲁棒性的对抗性训练方法。该方法中生成器用来从噪声特征中产生干净的特征表示,判别器用于区分干净信号和生成信号。2021年,Li等[26]提出了一种多鉴别器CycleGAN语音增强方法来提升语音识别性能,该方法不需要任何并行数据,通过设置能够比对不同频率区域的判别器和多个类似数据子集的生成器来优化噪声增强性能,从而提高自动语音识别的性能。

尽管效果提升显著,但生成对抗网络较难训练,通常利用频谱正则化等方法稳定优化过程,并且随着其他训练方法如自监督表示学习在低资源语言识别上展现出极其优异的性能,目前对其直接的使用逐渐减少,更多的是利用其进行领域自适应或者作为一种正则化手段结合其他方法一起使用。

2.2 自监督表示学习

自监督表示学习是无监督学习的一个分支,其本质是一种具有监督形式的非监督学习方法,即不借助任何标注数据,直接利用数据本身信息通过构造某种形式的辅助训练任务来学习对下游任务有价值信息的表示方法。由于其非常适合解决少样本环境下的模型冷启动问题,已经成为人工智能领域最热点的方向之一。自监督学习广义上可以分为生成式、对比式两类[27],而语音信号的自监督表示学习主要包括信号重建和对比预测两种,在语音识别、增强和处理等多个任务中获得显著效果。

基于信号重建的自监督学习方法属于生成式方法,即根据重建损失对语音信号的帧级基本单元信号进行重建。信号重建方法又可以细分为回归预测和掩蔽重建两个子类。比较典型的模型有自回归预测编码(autoregressive predictive coding, APC)[28]、矢量量化自回归预测编码(vector-quantized autoregressive predictive coding, VQAPC)[29]、重构变换器表示模型(transformer encoder representations from alteration, TERA)[30]、HuBERT[31]等。

语音处理中第二类自监督表示学习是对比式方法,其基本思想是通过自动构造相似实例(正例)和不相似实例(负例)训练一个表示模型,使得正例在编码器对应的投影空间中比较接近,反之负例距离较远,因此可以认为在编码器投影空间中学习到样本的本质特征,或者说找到了数据内在流形。而通常编码器训练准则为最小化噪声对比估计(noice-contrastive estimation, NCE)[32]或InfoNCE[33]。典型对比式自监督学习包括对比预测编码(contrastive predictive coding, CPC)、wav2vec[34]系列、XLST等。以wav2vec2.0为例,其对比损失 L_m 定义为

$$L_m = -\log \frac{\exp\left(\frac{sim(c_t, \boldsymbol{q}_t)}{k}\right)}{\sum_{\tilde{\boldsymbol{q}} \sim Q_t} \exp\left(\frac{sim(c_t, \tilde{\boldsymbol{q}})}{k}\right)}。 \tag{2}$$

式中:c_t 表示掩码时间步 t 时刻上下文编码器的输出。模型需要在一个有 K+1 个候选值的集合 $\tilde{\boldsymbol{q}} \in \boldsymbol{Q}_t$ 中找到正确的被掩码量化值 $\boldsymbol{q}_t$,同时降低它和其他量化值的距离。

虽然自监督模型在多个下游任务上取得了优异的性能,但是由于其在训练过程中采用的是无监督的范式,因此学习到的表示是通用的表示。对于少样本语音识别任务来说,这种表示存在相当大的容量冗余,在少样本的条件下模型难以有效去除掉这些冗余,并且模型的参数量巨大,在少样本条件下微调容易过拟合。因此对少样本语音识别任务来说,后续的研究主要集中在模型的轻量化调整、压缩学习到的表示中的无关容量等方面。

2.3　深度强化学习

强化学习是智能体(agent)与环境不断交互获得环境反馈进行学习的方法,本质上是一种“试错”学习,即在与环境的不断交互中寻找最优策略。其理论基础是马尔可夫决策过程(Markov decision process, MDP)。近几年来,强化学习与深度学习结合获得了巨大的发展,尤其是在 AlphaGo 凭借深度强化学习技术先后打败人类顶级围棋选手后,深度强化学习的应用越来越广泛。当前,深度强化学习技术在语音识别中应用较少,主要集中在两个方面[35]:一是利用策略梯度来进行模型优化提升,同时在目标函数不可导时可利用策略梯度算法无须计算回报函数导数的特点进行解决;二是利用强化学习无须数据标注的特点进行半监督训练解决少样本条件问题。

现有基于注意力机制编解码器结构的语音识别系统在训练时采用交叉熵训练准则和教师强制(teacher-forcing)训练方法,即解码器每一时刻输入强制约束为正确标注,而在测试时利用自回归方式进行解码后计算词错误率,即解码器每一时刻输入为上一时刻输出,使得训练和测试时解码器行为不一致。换而言之,解码器新预测单词在训练和测试时是从不同分布中推断,存在训练和测试失配问题。因此 Tjandra 等[35]提出采用强化学习训练方法,以输出序列与正确标注的编辑距离作为回报函数,采用策略梯度下降进行模型训练,在最大似然准则下错误率有明显下降。随后 Tjandra 等[36]进一步优化训练方法,探讨了两种不同单步符号级和序列级回报函数的性能,指出单步符号级回报函数可以获得更优结果。此外,为进行在线实时语音识别,Luo 等[37]提出一种类似于混合自回归变换器(hybrid autoregressive transducer, HAT)结构[38]的在线式序列到序列语音识别模型。

强化学习方法不需要监督信号,因此理论上也不需要标注数据,利用无标注数据就可以进行学习,但关键问题在于回报函数设定和识别结果的评价机制。由于强化学习本质是人机交互式学习,因此利用强化学习算法对语音识别系统进行半监督或非监督训练也成为一个重要发展方向。如 Kala 等[39]提出一种利用无标注数据、以人机交互式学习提升语音识别系统性能的方法;Chung 等[40]提出一种强化学习半监督训练方法,即对标注数据采用交叉熵准则进行训练;Radzikowski 等[41]提出将对偶学习和强化学习相结合的方法进行非母语说话人语音识别系统的半监督训练。最近由于 ChatGPT 的优异表现,基于人类反馈的强化学习引起了前所未有的关注,后期如何将相关的研究引入到语音识别中,并根据领域特点进行调整是可期望的一个研究方向。

2.4　元学习

由于深度学习方法性能极其受标注数据规模的影响,因此迫切需要一种新型学习方式来实现少样本甚至是零样本的学习。神经网络元学习作为推动当代深度学习行业前沿的潜在强大驱动力,导致了近期研究的爆炸式增长。业界期望通过元学习来解决当代深度学习中的许多瓶颈问题。

2019 年起研究者开始将元学习中的 MAML 等算法应用到低资源语音识别领域。在语音识别中,经常将不同语言作为不同的任务,假设元任务集合 $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K\}$, MAML 或者 Reptile 利用任务集合来获得好的初始化参数 θ。对于某个任务数据 $\mathcal{D}_t$, 可以得到更新后的参数 θ_t, 那么模型无关的元学习方法(如 MAML 或者 Reptile)可以表示为

$$\theta_t = Learn(\mathcal{D}_t, \theta) = Learn(\mathcal{D}_t, MetaLearn(\mathcal{D}))。\quad (3)$$

从式(3)可以看出,这是一个双优化过程:内优化过程是 $Learn(\mathcal{D}_t, \theta)$;外优化过程是 $MetaLearn(\mathcal{D})$。

给定初始化参数 θ^0 和任务数据 $\mathcal{D}_t$, 则任务特有的学习表示为

$$\theta'_t = Learn(\mathcal{D}_t, \theta^0) = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\mathcal{D}_t}(\theta)。\quad (4)$$

假设任务数据 $\mathcal{D}_t$ 进一步划分为支持集 $\mathcal{D}_t^{\mathrm{tr}}$ 和查询集 $\mathcal{D}_t^{\mathrm{te}}$, 即 $\mathcal{D}_t = \{\mathcal{D}_t^{\mathrm{tr}}, \mathcal{D}_t^{\mathrm{te}}\}$, 则有

$$\theta_{Meta} = MetaLearn(\mathcal{D}, \{\theta'_t\}_{t=1}^{K}) = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\mathcal{D}}^{Meta}(\theta);\quad (5)$$

$$\mathcal{L}_{\mathcal{D}}^{Meta}(\theta) = E_{\tau_k \sim p(\tau)} E_{\mathcal{D}_t^{\mathrm{tr}}, \mathcal{D}_t^{\mathrm{te}}}\, \mathcal{L}_{\mathcal{D}_t^{\mathrm{te}}}(\theta)。\quad (6)$$

王璐等[42]在 TIMIT 和自建佤语语料库上利用经典 MAML 算法和 Reptile 算法,探讨并验证了元学

习方法有助于解决小规模孤立词语音识别。后期侯俊龙等[43]将元度量学习也用于低资源孤立词识别,但其孤立词应用场景设定过于简单缺乏实用性。2018年Klejch等[44]提出采用元学习方法来进行说话人自适应来解决训练和测试条件失配问题,研究结果表明元学习者可以学习执行有监督和无监督说话者适应,并且在适应具有1.5 M个参数的DNN声学模型时,优于当前性能比较好的隐含单元贡献度学习(learning hidden unit contribution, LHUC)深层网络自适应方法。Hsu等[45]提出端到端框架的元语音识别(MetaASR)方法,即将不同语言的ASR认为是不同的任务,并且利用MAML算法进行模型初始化训练,使用6种语言作为预训练任务,4种语言作为目标任务。结果表明,MetaASR显著优于最先进的多任务预训练方法。2020年,Xiao等[46]借鉴MetaASR方法构造多语言元学习语音识别(multilingual meta-learning ASR, MML-ASR)方法,将每个源语言ASR问题分解成许多小的ASR识别任务,然后利用元学习对不同源语言所有任务的模型初始化,以快速适应未知目标语言。但由于不同语言的数据规模和其发音系统建模难度差异性很大,导致了元学习子任务数量和任务难度不平衡问题,从而导致了多语言元学习方法的失败,为此该团队又提出元对抗采样学习方法(adversarial meta sampling, AMS)来改善不同语言子任务的非均衡问题[46]。口音的较大可变性和复杂性为语音识别系统带来重大挑战,2020年,Winata等[47]在MAML算法基础上提出口音无关元学习方法来快速适应未知英语口音,在混合区域和跨区域口音任务中该方法均优于联合训练方法。同年,Winata等[48]又提出基于MAML的元迁移学习方法对混合语种的连续语音进行识别,结果优于同等情况下的其他方法。

元学习本质上是一种更通用的模式,其核心在于元知识(meta-knowledge)的表征和获取。通过在任务上学习,具有非常强大的表征能力,因此对新任务的泛化能力更强。目前在连续语音识别上应用的元学习算法主要是基于梯度的MAML、Reptile、ANIL等算法,这些算法都是与模型无关的、只改变模型的训练方式。元学习将语言作为任务,通过“先适应再学习”的学习范式,能够获得对新任务快速适应的能力。

以上4类高级深度学习技术各有特点,每类技术的代表方法、优缺点以及适用场景如表1所示。

表1 高级深度学习技术方法特点对比

Table 1 Comparison of advanced deep learning techniques and methods

技术类别	代表方法	优点	适应场景	缺点
生成对抗网络	FGSM	基于参数动态生成	噪声和变化信道	样本受模型影响大
自监督表示学习	Wav2Ve	少量数据实现高性能	大量无标注数据	需要消耗较多计算资源
深度强化学习	HAT	实时性高	在线语音识别	需要设计评价机制
元学习	AMS	较快适应新语言	任务难度不平衡的多语言预训练	需要额外训练一个采样网络

3 存在的问题和挑战

尽管研究人员在低资源语音识别方面有一定进展,但从认知角度来看,采用的方法与人的快速学习能力相差甚远;而且在训练数据规模方面,没有考虑到更小的规模,因此需要对语音识别采用更先进的理论、对低资源标注数据获取条件更为受限的情况开展研究。目前面临的主要问题包括以下几点。

3.1 不同学习范式的组合互补有限

低资源少样本学习的总误差受假设空间$\mathcal{H}$和训练样本数的影响,可以从训练数据、确定假设空间$\mathcal{H}$的模型和搜索最佳假设算法3个方面来克服。其中,生成对抗网络、自监督表示学习、深度强化学习和元学习从不同角度解决少样本带来的影响,虽然获得了可喜成果,但这些方法都是从自身机理出发自成一脉,相互之间虽有部分融合,但互补性并没有充分发挥,尤其是缺乏体系化的组合互补策略,因此需要从数据、模型和算法等更高层面将这些方法进行有机组合,以提升连续语音识别系统的整体性能水平。

3.2 多语言联合学习对数据和任务均衡问题考虑不足

由于很多语言的单一语种数据都很匮乏,而借鉴不同语言之间的共性信息进行多语言联合学习已经成为提升少样本语言识别性能的一种有效方式。无论是自监督表示学习、元学习或者生成对抗网络,在进行多语言联合处理时都对不同语言没有区别对待。但实际过程中,即使是相同方法处理不同语言的数据时性能都会有一定差异,或者说每种语言都有不同的难度系数。以元学习模型构建为例,初始化元学习方法由于上述不均衡性会导致其初始化模型离大规模、易训练收敛语言比较近,离目标语言最优模型较远,会导致后期优化进程缓慢且容易陷入局部最优。因此在采用不同方法进行模型建模时,面临多种“不均衡”问题,如何克服这种不均衡性、减少不均衡性对系统性能的影响是值得思考的问题。

3.3　模型轻量化部署需更深入研究

深度学习由于其强大的复杂模式表示能力使得其在语音识别等诸多领域获得了突飞猛进的发展,但由于其模型参数规模大、计算复杂度和空间复杂度高,无法有效应用于轻量级资源受限的便携设备中。而语音识别是人机交互的重要一环,人机交互场景更需要小型化、微型化设备的应用,因此如何在低资源少样本条件下进行模型部署也成为一个重要问题。

深度学习技术正朝着两极发展:一是深度学习领域研究人员致力于研发更深、更大的模型,达到更高的精度和准确度,如 speech-transformer 模型层数和参数规模都很大,且这种大模型也开始向微观世界发展,如深度学习用于蛋白质合成、分子发现等领域;二是深度学习自身朝着小型化发展,很多智能化应用场景的搭载平台受体积、功耗等因素影响,因此一些研究学者致力于对深度学习模型进行压缩以便部署在小型平台。而作为深度学习应用的重要领域,语音识别也同样遵循上述两极化发展的脉络。目前模型压缩技术主要包括浅层压缩和深层压缩两大类,浅层压缩主要通过裁剪和知识整流来实现,而深层压缩通过量化、轻量级网络和结构搜索来实现。但当模型压缩与低资源少样本同时出现时,其性能更难以保证,因此未来需要对模型轻量化技术进行更加深入的研究,以便在低资源少样本条件下,轻量化模型可取得期望的性能。

4　结语

事实上,众多研究证实:虽然低资源少样本语音识别在标注数据获取、高层次语义表征、紧致模型的有效表征训练等方面存在诸多困难,但从少样本学习误差理论分析可知,少样本训练识别仍然可以通过先验信息引入、假设空间约束条件设定等方式优化提升。现有的自监督表示学习、元学习等高级深度学习技术都在低资源少样本语音识别领域展现了优越的性能。未来这些高级深度学习技术的体系化的组合互补策略、克服语言之间的不均衡性以及对深度模型的压缩与轻量化部署等方面都是值得进一步研究的方向。

参考文献:

[1] 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. New York: ACM, 2006: 369-376.

[2] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2013: 6645-6649.

[3] 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). Piscataway: IEEE, 2016: 4960-4964.

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

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-03-10].https://arxiv.org/abs/1810.04805.

[6] GALES M J F, KNILL K M, RAGNI A, et al. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED[C]//The 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages. St. Petersburg: RFBR, 2014: 16-23.

[7] 赵淑芳, 董小雨. 基于改进的 LSTM 深度神经网络语音识别研究[J]. 郑州大学学报(工学版), 2018, 39(5): 63-67.
ZHAO S F, DONG X Y. Research on speech recognition based on improved LSTM deep neural network[J]. Journal of Zhengzhou University (Engineering Science), 2018, 39(5): 63-67.

[8] THOMAS S, GANAPATHY S, HERMANSKY H. Multilingual MLP features for low-resource LVCSR systems[C]// 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Pisca-taway: IEEE, 2012: 4269-4272.

[9] POVEY D, BURGET L, AGARWAL M, et al. The subspace Gaussian mixture model: a structured model for speech recognition[J]. Computer Speech & Language, 2011, 25(2): 404-439.

[10] IMSENG D, BOURLARD H, GARNER P N. Using KL-divergence and multilingual information to improve ASR for under-resourced languages[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2012: 4869-4872.

[11] MOHAMED A R, DAHL G E, HINTON G. Acoustic modeling using deep belief networks[J]. IEEE

Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

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

[13] 薛均晓, 黄世博, 王亚博, 等. 基于时空特征的语音情感识别模型 TSTNet[J]. 郑州大学学报(工学版), 2021, 42(6): 28-33.
XUE J X, HUANG S B, WANG Y B, et al. Speech emotion recognition TSTNet based on spatial-temporal features[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(6): 28-33.

[14] POVEY D, PEDDINTI V, GALVEZ D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]// Interspeech 2016. San Francisco: ISCA, 2016: 2751-2755.

[15] JAITLY N, HINTON E. Vocal tract length perturbation (VTLP) improves speech recognition[C]// Proceedings of the Workshop on Deep Learning for Audio, Speech and Language. Atlanta: ICML, 2013:1-5.

[16] KO T, PEDDINTI V, POVEY D, et al. Audio augmentation for speech recognition[C]// Interspeech 2015. Dresden: ISCA, 2015: 3586-3589.

[17] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[EB/OL]. (2019-04-18)[2023-03-10]. https://arxiv.org/abs/1904.08779.

[18] KHARITONOV E, RIVIÈRE M, SYNNAEVE G, et al. Data augmenting contrastive learning of speech representations in the time domain[C]//2021 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE, 2021: 215-222.

[19] XIE Q Z, LUONG M T, HOVY E, et al. Self-training with noisy student improves ImageNet classification[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 10684-10695.

[20] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. (2014-06-10)[2023-03-10]. https://arxiv.org/abs/1406.2661.

[21] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
WANG K F, GOU C, DUAN Y J, et al. Generative adversarial networks: the state of the art and beyond[J]. Acta Automatica Sinica, 2017, 43(3): 321-332.

[22] QIAN Y M, HU H, TAN T. Data augmentation using generative adversarial networks for robust speech recognition[J]. Speech Communication, 2019, 114: 1-9.

[23] SUN S N, YEH C F, OSTENDORF M, et al. Training augmentation with adversarial examples for robust speech recognition[EB/OL].(2018-06-07)[2023-03-10].https://arxiv.org/abs/1806.02782.

[24] SHINOHARA Y. Adversarial multi-task learning of deep neural networks for robust speech recognition[C]// Interspeech 2016. San Francisco: ISCA, 2016: 2369-2372.

[25] LIU B, NIE S, ZHANG Y P, et al. Boosting noise robustness of acoustic model via deep adversarial training[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 5034-5038.

[26] LI C Y, VU N T. Improving speech recognition on noisy speech via speech enhancement with multi-discriminators CycleGAN [C]// 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE, 2022: 830-836.

[27] 屈丹, 张文林, 杨绪魁. 实用深度学习基础[M]. 北京: 清华大学出版社, 2022.
QU D, ZHANG W L, YANG X K. Practical deep learning foundation[M]. Beijing: Tsinghua University Press, 2022.

[28] CHUNG Y A, HSU W N, TANG H, et al. An unsupervised autoregressive model for speech representation learning[C]// Interspeech 2019. Graz: ISCA, 2019: 146-150.

[29] CHUNG Y A, TANG H, GLASS J. Vector-quantized autoregressive predictive coding[C]// Interspeech 2020. Shanghai: ISCA, 2020: 3760-3764.

[30] LIU A T, LI S W, LEE H Y. TERA: self-supervised learning of transformer encoder representation for speech [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2351-2366.

[31] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021, 29: 3451-3460.

[32] GUTMANN M, HYVÄRINEN A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models [J]. Journal of Machine Learning Research, 2010, 9: 297-304.

[33] OORD A V D, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding [EB/OL]. (2019-01-22)[2023-03-10]. https://arxiv.org/abs/1807.03748.

[34] SCHNEIDER S, BAEVSKI A, COLLOBERT R, et al. Wav2vec: unsupervised pre-training for speech recognition [C]// Interspeech 2019. Graz: ISCA, 2019: 3465-3469.

[35] TJANDRA A, SAKTI S, NAKAMURA S. Sequence-to-sequence ASR optimization via reinforcement learning[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 5829-5833.

[36] TJANDRA A, SAKTI S, NAKAMURA S. End-to-end speech recognition sequence training with reinforcement learning[J]. IEEE Access, 2019, 7: 79758-79769.

[37] LUO Y P, CHIU C C, JAITLY N, et al. Learning online alignments with continuous rewards policy gradient[C]// 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2017: 2801-2805.

[38] VARIANI E, RYBACH D, ALLAUZEN C, et al. Hybrid autoregressive transducer (HAT)[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 6139-6143.

[39] KALA T K, SHINOZAKI T. Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 5759-5763.

[40] CHUNG H, JEON H B, PARK J G. Semi-supervised training for sequence-to-sequence speech recognition using reinforcement learning[C]// 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2020: 1-6.

[41] RADZIKOWSKI K, NOWAK R, WANG L, et al. Dual supervised learning for non-native speech recognition[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2019, 2019(1): 1-10.

[42] 王璐, 潘文林. 基于元学习的语音识别探究[J]. 云南民族大学学报(自然科学版), 2019, 28(5): 510-516.
WANG L, PAN W L. Speech recognition based on meta-learning[J]. Journal of Yunnan Minzu University (Natural Sciences Edition), 2019, 28(5): 510-516.

[43] 侯俊龙, 潘文林. 基于元度量学习的低资源语音识别[J]. 云南民族大学学报(自然科学版), 2021, 30(3): 272-278.
HOU J L, PAN W L. Low-resource speech recognition based on meta-metric learning[J]. Journal of Yunnan Minzu University (Natural Sciences Edition), 2021, 30(3): 272-278.

[44] KLEJCH O, FAINBERG J, BELL P. Learning to adapt: a meta-learning approach for speaker adaptation[C]// Interspeech 2018. Hyderabad: ISCA, 2018: 867-871.

[45] HSU J Y, CHEN Y J, LEE H Y. Meta learning for end-to-end low-resource speech recognition[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 7844-7848.

[46] XIAO Y B, GONG K, ZHOU P, et al. Adversarial meta sampling for multilingual low-resource speech recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14112-14120.

[47] WINATA G I, CAHYAWIJAYA S, LIU Z H, et al. Learning fast adaptation on cross-accented speech recognition[C]// Interspeech 2020. Shanghai: ISCA, 2020: 1276-1280.

[48] WINATA G I, CAHYAWIJAYA S, LIN Z J, et al. Meta-transfer learning for code-switched speech recognition[EB/OL]. (2020-03-04)[2023-03-10]. https://arxiv.org/abs/2003.01901.

Overview of Recent Progress in Low-resource Few-shot Continuous Speech Recognition

QU Dan, YANG Xukui, YAN Honggang, CHEN Yaqi, NIU Tong

(School of Information System Engineering, Strategic Support Force Information Engineering University, Zhengzhou 450001, China)

Abstract: Low-resource few-shot speech recognition is an urgent technical demand faced by the speech recognition industry. The framework technology for few-shot speech recognition was first briefly discussed in this study. The research progress of several important low resource speech technologies, including feature extraction, acoustic model, and resource expansion, was then highlighted. The latest advancements in deep learning technologies, such as generative adversarial networks, self-supervised representation learning, deep reinforcement learning, and meta-learning, were then focused on how to address few-shot speech recognition on the basis of the development of continuous speech recognition framework technology. On that basis, the problems of limited complementarity, unbalanced task and model deployment faced by this technology were analyzed for the subsequent development. Finally, a summary and prospects of few-shot continuous speech recognition were given.

Keywords: low-resource few-shot; continuous speech recognition; generative adversarial networks; self-supervised representation learning; deep reinforcement learning; meta-learning