# 基于样本类不确定性抽样的端到端语音关键词检测训练方法

贺前华,陈永强,郑若伟,黄金鑫
(华南理工大学电子与信息学院,广东广州 510641)

**摘 要:** 当前语音关键词检测主流技术为端到端的深度学习方法,研究重点为网络结构优化、建模单元选取及搜索策略等,并取得较快进展,但对模型训练效率的关注相对较少. 本文针对深度学习模型训练效率问题,提出了一种样本类不确定性抽样(Class Uncertainty Sampling,CUS)的样本应用策略加速收敛进程. 其核心思想是在模型训练中后期,利用网络的前向输出层对样本评价信息进行样本类不确定性度量,并转化成样本选用概率,随机抽取训练样本子集用于后续训练. 由于简单样本的类确定度高,它们参与后续训练的概率降低,但不影响模型的区分能力,增强对判决边界样本的关注,达到提高模型训练效率的目标. 基于AISHELL-1普通话数据集的实验结果表明,相对常规训练策略,平均训练时长缩短60%,收敛时长缩短47.5%. 虚警率(False Alarm Rate,FAR)为0.5 FP/h时,该方法的错误拒绝率(False Reject Rate,FRR)从4.75%降至3.65%,相对下降30.1%,最大关键词加权值(Maximum Term Weighted Value,MTWV)由0.837 4升至0.853 1. 通过分析错标样本参与训练的行为,证实了该方法具有屏蔽掉大部分错误标注样本的能力,减少错标样本对训练的损害. 基于大规模AISHELL-2普通话数据集的实验进一步证实了提出方法的有效性.

**关键词:** 语音关键词检测;深度学习;端到端;类不确定性抽样
**基金项目:** 广东省科技计划项目(No.2023A0505050116,No.2022A1515011687);国家自然科学基金(No.62371195)
**中图分类号:** TN912;TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3482-11
**电子学报URL:**http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240048

# End-to-End Speech Keyword Spotting Training Method Based on Sample's Class Uncertainty

HE Qian-hua, CHEN Yong-qiang, ZHENG Ruo-wei, HUANG Jin-xin
(*School of Electronic and Information Engineering*, *South China University of Technology*, *Guangzhou*, *Guangdong* 510641, *China*)

**Abstract:** End-to-end deep learning is the main technology for speech keyword spotting. The research focused on exploring better network structures, modeling units, and search strategies, and have made a lot of progress. However, less attention is paid on training efficiency. In this paper, a novel class uncertainty sampling (CUS) strategy is proposed to select effective samples for each training epoch. Since only a subset is used, much training time is saved. The core idea of CUS is measuring the class uncertainty of samples with the forward information of the output layer during the middle and late training stages, and samples are selected at a probability of their class uncertainty. Therefore more attention is paid to samples nearing the decision boundary, which are prone to missed detection or false alarm. Furthermore, the proposed method could shield the interference of label error samples. Experimental results on the AISHELL-1 Mandarin dataset showed that fast convergence and better training performance were achieved. Against the conventional training strategy, the average training time and the average converging time was relatively shortened by 60% and 47.5%, respectively. At 0.5 FP/h false accept rate (FAR), the false reject rate (FRR) was reduced from 4.75% to 3.65%, a relative reduction of 30.1%, and the maximum term weighted value (MTWV) was increased from 0.837 4 to 0.853 1. Moreover, it was experimentally verified that the method could shield most of the mislabeled samples. This conclusion was confirmed with the experiments on the large-scale AISHELL-2 Mandarin dataset.

**Key words:** speech keyword spotting; deep learning; end-to-end; class uncertainty sampling

收稿日期:2024-01-10;修回日期:2024-03-08;责任编辑:赵克

Foundation Item(s): Guangdong Science Foundation (No.2023A0505050116, No.2022A1515011687); National Nature Science Foundation of China (No.62371195)

## 1 引言

语音关键词检测(KeyWord Spotting,KWS)已广泛应用于设备唤醒[1]、语音检索[2]、智能家居[3]以及物联网[4]等领域. 尽管有多种技术路线[5],但随着深度学习的发展,端到端(End to End,E2E)语音关键词检测方法可以直接学习语音特征与目标输出的映射关系,不需要对标注数据进行强制预对齐[6],大大简化训练流程和系统结构的复杂性,成为当前研究热点[7,8].

端到端语音关键词检测方法分为基于注意力机制编码-解码器(Attention-based Encoder Decoder,AED)[9]和基于连接时序分类器(Connectionist Temporal Classification,CTC)[10]2种类型. Shan等人[11]最先提出利用注意力机制构建关键词检测模型,在取得高检测性能的同时简化语音关键词检测系统. Higuchil等人[12]提出结合跨注意力机制的多任务学习方案,显著提升关键词检测性能. Bai等人[13]则提出使用循环神经网络(Recurrent Neural Network,RNN)和CTC训练端到端中文关键词检测模型,随后研究者们进一步对建模单元、网络拓扑结构和损失函数等进行探索. Yan等人[14]提出使用卷积循环神经网络(Convolutional Recurrent Neural Network,CRNN)结构和有调音节来构建关键词检测模型,进一步提升性能. Lan等人[15]提出了加权CTC损失函数缓解样本不平衡问题. 在这些方法中,通过利用CNN捕获局部信息能力和RNN的时序记忆能力,以CRNN结构和有调音节作为建模单元的端到端语音关键词检测方法,取得更优的检测性能. 本文主要基于CTC的端到端语音关键词检测开展研究.

深度学习是1个循环迭代的过程,依赖于大算力和大样本,目前常用的端到端随机梯度下降法(Stochastic Gradient Descent,SGD)缺乏对模型训练效率的关注. 在常规模型训练策略中,样本以一样的权重参与训练,随着训练的进行越来越多的容易样本(经过简单训练就能被模型识别的样本,一般处于类样本空间的中心区域)被学好,对梯度的贡献和模型参数调整的影响也越来越小,重复训练简单样本将导致大量无效计算,造成计算资源和能源浪费,因此,深度训练成本问题成为关注的热点问题[16]. 除了诸如增加硬件资源[17]、参数初始化、批标准化、改进模型结构、自适应学习率优化算法等通用的提高模型训练速度的技巧外,还可以通过优化模型参数更新规则[18],设计更合理的样本应用策略加速模型训练. Zhang等人[19]提出在训练过程中选择尽量不相似的样本,构建梯度噪声方差更低的批,减轻随机梯度噪声对收敛速度的影响,加快模型收敛. 与其思路类似,Peng等人[20]则提出使用典型采样的方式构建批来提高模型训练速度. Jiang等人[21]提出用最大损失采样的方式进行选择性梯度反向传播,降低常规训练反向传播的计算次数,加速模型的训练过程. 文献[16]则采用梯度估计的思路降低梯度计算成本. 迁移学习是提高模型训练速度的另外一种选择[22]. 然而,此类样本应用策略方法主要以提升训练速度为目的,忽略了易造成漏检和虚警这类样本的特殊性,在提升模型训练速度时会产生性能明显下降. 尽管课程学习(Curriculum Learning,CL)[23]专注于样本应用策略,但着重一般性理论和原则,没有考虑语音关键词检测任务的特点及分类决策过程.

在语音关键词检测任务中,易造成漏检和虚警的样本主要是靠近决策边界的样本,相邻类的边界样本具有较高的类相似度. 另一方面,相对容易样本,它们的数量要少得多,为了提高模型对这些样本的检测能力,学习过程中需要给予更多关注,挖掘其中的隐性类区分特征. 而样本均权训练将导致这些关键信息淹没在大量简单样本的信息中. 本文综合考虑深度网络在训练过程中的成熟度、每个样本在训练过程中的动态变化及决策机制,将深度分类器的训练过程分为2个阶段. 训练开始时,深度网络不具备任何判决能力,利用全体样本获得类初始边界,当深度网络具备一定类区分能力后,进入相邻类边界优化阶段. 一般而言,优化阶段的训练轮次远比初始边界的建立要多(参见文献[14]),因此提出了一种类不确定性抽样(Class Uncertainty Sampling,CUS)的样本应用策略,即在模型训练中后期,利用模型的初步区分能力对语音样本进行类不确定性度量,并以此定义样本参与下1轮训练的概率,使类不确定度高的样本参与概率高,用概率抽样构建下1轮训练的样本集. 从整体训练过程来看,该机制使类边缘样本有更多机会参与后续训练,弥补了边界样本与类中心样本不平衡造成的训练偏置问题,同时减少类确定度高样本的重复训练,具有屏蔽类标注错误等低质量样本干扰的能力,提高模型学习效率.

本文符合CL的基本思想:选择哪些样本用何种顺序提供给学习系统. 但具体实施方法不同:(1)样本的组织是依据模型对样本的类隶属度度量的相对值,而前面提到的方法均采用损失值(绝对值),使模型获得更好的判决能力;(2)样本的组织是依概率,而不是基于门限,从而缓解训练遗忘问题,保证了模型的泛化能力;(3)使用训练样本量是逐渐减少的,降低大量无效

计算,而CL的先易后难策略是将难度大的样本逐渐加到训练集中[23],使训练过程中样本逐渐增加.

## 2 基于样本类不确定性的样本筛选方法

本节首先介绍CUS算法框架,然后结合CTC端到端语音关键词检测方法,对类不确定性度量给出具体定义,并进一步定义参与概率和抽样开启条件,对超参数设计进行详细阐述.

### 2.1 CUS算法框架

图1为CUS算法框架,分为常规训练和选择样本训练2个阶段. 首先采用常规训练,样本集$D$中所有样本参与训练,使模型快速学习类样本的共性信息及类间显著区分信息,如汉字的基本发音特征,建立基本的类间决策边界面. 在训练中后期,模型训练的重点放在类区分性不显著的样本区分信息的利用上,挖掘利用声学相似度高的不同类样本中的隐性区分性信息,实现对决策边界面的微调,获得更优决策界面. 为减少额外计算量,本文方法利用模型输出层样本的类信息计算每个样本的类不确定度,依此定义样本参与下1轮训练的概率,该概率与样本的类不确定度成反比,然后依照概率构建训练子集$D'\subseteq D$. 从训练过程来看,类稳定样本参与度低,加大对易造成漏检和虚警这一类边界样本的关注,提高模型整体训练效率和性能. 采用概率样本选择方式,有效避免神经网络遗忘问题,当1个样本的类确定性下降后,其参与下1轮训练的概率就提高了. 算法1为CUS算法的详细描述,其中,$(\boldsymbol{x}_n,\boldsymbol{y}_n)$为训练集中第$n$个音频样本及其标注,$X^k$为1轮训练中第$k$个批次的样本集,而$\boldsymbol{X}_m^k$为$X^k$中的第$m$个样本.

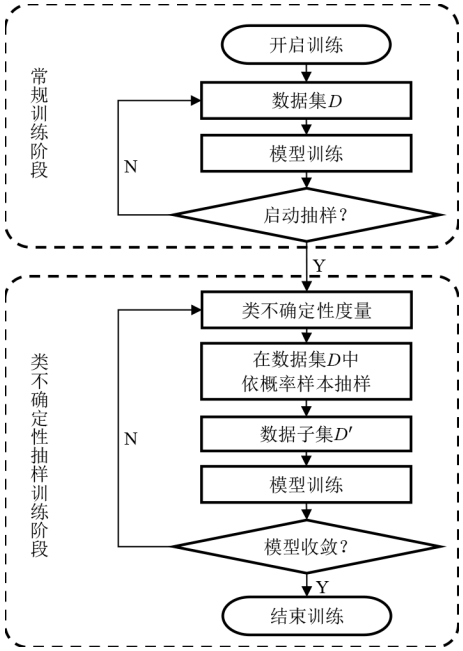



图1 类不确定性抽样的算法框架

本文方法仅利用当前模型对样本的类评价信息对样本进行类不确定度评价,动态构建下1轮训练的样本集,对深度模型和模型训练方法没有约束,具有即插即用的特点.

**算法1 基于类不确定性抽样的训练过程**

**输入**: 训练集$D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$,学习率$\eta$,批大小$M$,未经训练的模型参数矩阵$\boldsymbol{\theta}$,抽样开启条件

初始化: $D'\leftarrow D$,$p(\boldsymbol{x}_i)=1$
while(模型未收敛) do
  将$D'$随机打乱并将其划分为$K=\text{ceil}(\frac{|D'|}{M})$个批次
  for $k=1$ to $K$ do
    前向传播:output$(X^k)\leftarrow$Model$(X^k)$
    损失计算:$L(\boldsymbol{\theta}_k)\leftarrow$Loss(output$(X^k)$)
    参数更新:$\boldsymbol{\theta}_k\leftarrow\boldsymbol{\theta}_{k-1}-\eta\nabla_{\theta}L(\boldsymbol{\theta}_k)$//对当前批次中样本进行类不确定性度量
    if 达到抽样开启条件 then
      for $m$=1 to $M$ do
        CU$(\boldsymbol{X}_m^k)\leftarrow$ClassUncertainty(output$(\boldsymbol{X}_m^k)$)
        $p(\boldsymbol{X}_m^k)\leftarrow$CalcProbability(CU$(\boldsymbol{X}_m^k)$)
      end for
    end if
  end for //在整个训练集中依照概率抽样构建训练子集
  if 达到抽样开启条件 then
    $D'\leftarrow\{\}$
    for $i=1$ to $N$ do
      randnum$\leftarrow$rand$(0,1)$
      if $p(\boldsymbol{x}_i)>$randnum then
        $D'$.append$(\boldsymbol{x}_i)$
      end if
    end for
  end if
end while

**输出**: 经过训练优化的模型参数$\boldsymbol{\theta}$

### 2.2 样本类不确定性度量

基于信息论知识,本文定义语音关键词样本的类不确定性为语音关键词检测模型输出结果的不确定程度. 当结果很明确时,样本的类不确定性低,隐藏在不同类样本之间的区分信息是否再挖掘利用不影响该样本的类判决;而当结果不够明确时,样本的类不确定性高,处于决策边界附近,需要进一步挖掘样本区分于其他类样本的信息,提高类判决的可靠性. 为有效度量样本的类不确定性,对于样本$\boldsymbol{x}_i$,定义以下参量:

(1)目标类得分$S_{\text{t}}(\boldsymbol{x}_i)$:在模型输出后验概率矩阵中进行关键词搜索,得到目标关键词得分.

(2)竞争类得分$S_c(\boldsymbol{x}_i)$:除目标类得分外的最高得分,视为目标的竞争者.

当模型具有一定决策能力后,根据样本的目标类得分和竞争类得分,将样本$\boldsymbol{x}_i$分为下述3种类型:

(1)简单样本:类特征显著的样本,样本的目标类得分$S_t(\boldsymbol{x}_i)$远高于竞争类得分$S_c(\boldsymbol{x}_i)$,即$S_t(\boldsymbol{x}_i) \gg S_c(\boldsymbol{x}_i)$. 此时样本属于正确检测,结果确定度高,样本类不确定性低. 表明样本已能被模型稳定识别,过多参与后续训练并不能带来性能提升,且导致大量无效计算,甚至可能导致过度训练的问题;另一方面,适当减少这类样本的参与度,$S_t(\boldsymbol{x}_i) > S_c(\boldsymbol{x}_i)$是大概率事件,因此不会导致其识别性能的下降. 因这类样本在样本集中占绝大多数,后续训练中应减少这类样本的参与,可减少大量无效计算.

(2)低质量样本:样本的目标类得分$S_t(\boldsymbol{x}_i)$远低于竞争类得分$S_c(\boldsymbol{x}_i)$,即$S_t(\boldsymbol{x}_i) \ll S_c(\boldsymbol{x}_i)$. 此时样本检测错误,但结果的确定度高,样本类不确定性同样很低. 这类样本大多数属于标注错误,此类样本在训练过程中会产生过大的梯度,造成模型不易收敛甚至损坏模型的性能,同样应减少这类样本参与后续训练.

(3)边界样本:样本的目标类得分$S_t(\boldsymbol{x}_i)$与竞争类得分$S_c(\boldsymbol{x}_i)$相差不大,样本的类不确定性高. 在正确区分下,结果可靠程度不高,在错误区分情况下,稍微努力有可能变为正确区分. 当前模型对这类样本的区分能力有待提升,需要着重训练,挖掘这类样本的非显著区分性信息. 这类样本大概率处于决策边界附近. 现实中,语言单元的声学特征空间完全分离的概率极小,因此边界样本的利用也应有一定限度,目标限于寻找错分率较低的优化边界,而不能区分所有样本的最优边界.

上述分析表明,$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$越大,样本类不确定性越低,$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$越小,样本类不确定性越高,因此样本$\boldsymbol{x}_i$的类不确定性与$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$存在反比关系,如式(1)所示:

$$\text{Class Uncertainty}(\boldsymbol{x}_i) \propto \frac{1}{\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|} \tag{1}$$

### 2.3　目标类得分和竞争类得分

语音样本可以分成包含关键词的样本和不包含关键词的样本. 对于包含目标关键词$\boldsymbol{w}_k \in W$的样本,$W=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_K\}$为预定义的$K$个关键词,在Top-1决策机制下,有以下2种情况:

(1)若目标关键词$\boldsymbol{w}_k$的得分$S_t(\boldsymbol{x}_i) > S_c(\boldsymbol{x}_i)$值,为正检,否则为漏检.

(2)若有虚警词$\boldsymbol{w}_n(n \neq k)$的得分$S_c(\boldsymbol{x}_i) > S_t(\boldsymbol{x}_i)$,则为虚警,否则无虚警.

1个语音样本,可能检出多个正检或虚警,也可能完全没有检出(漏检),这2种情况如下:

(1)检出1个或多个关键词. 图2为1个正检和1个虚警的情况. 首先考虑正检区域,此时目标类得分$S_t(\boldsymbol{x}_i)$直接定义为目标关键词$\boldsymbol{w}_k$的概率,在该区域除目标关键词$\boldsymbol{w}_k$外的其他结果作为竞争对象,除目标关键词$\boldsymbol{w}_k$外的最大概率作为竞争类得分$S_c(\boldsymbol{x}_i)$;对于虚警区域,此时竞争类得分$S_c(\boldsymbol{x}_i)$定义为虚警词$\boldsymbol{w}_n$的概率,在该区域的目标对象应该为非关键词,因此,在该音段获得非关键词的概率作为目标类得分$S_t(\boldsymbol{x}_i)$;最后比较2个区域的$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$,较小者意味目标与竞争比较接近,其类不确定性高. 若出现多个区域,选择$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$最小者用于类不确定性度量;若只出现单个区域则可直接进行度量.

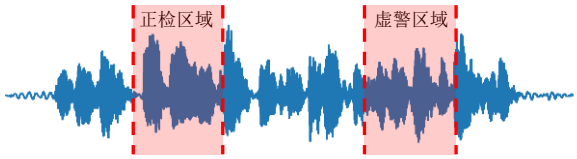



图2　检出多个关键词时的情况

(2)无关键词检出. 此时目标对象为目标关键词$\boldsymbol{w}_k$,意味着出现漏检,但仍然可以在输出矩阵中找到目标关键词$\boldsymbol{w}_k$最大概率,视为此音段最有可能出现目标关键词的区域. 因此,目标类得分$S_t(\boldsymbol{x}_i)$定义为目标关键词$\boldsymbol{w}_k$的最大概率,同时在该音段的非关键词概率作为竞争类得分$S_c(\boldsymbol{x}_i)$. 在此漏检区域,$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$之间的差值相对于其他正检区域会较小,此类漏检样本的类不确定性相对较高.

对不包含关键词的样本,所有音段的目标对象均为非关键词,而竞争对象是可能造成虚警的任意预定义关键词$\boldsymbol{w}_j, j \in \{1,2,\cdots,K\}$. 则将输出结果中最大的预定义关键词$\boldsymbol{w}_j$的概率作为竞争类得分$S_c(\boldsymbol{x}_i)$,同时在对应音段获得非关键词的概率作为目标类得分$S_t(\boldsymbol{x}_i)$. 当没有出现虚警时,非关键词的概率很大,目标类得分$S_t(\boldsymbol{x}_i)$远大于竞争类得分$S_c(\boldsymbol{x}_i)$,此类样本的类不确定性较小;当出现虚警时,在样本标注正确的前提下,竞争类得分$S_c(\boldsymbol{x}_i)$会相对变大,$\left|S_t(\boldsymbol{x}_i)-S_c(\boldsymbol{x}_i)\right|$变小,虚警样本具有更高的类不确定性,可以更多关注虚警样本.

### 2.4　样本参与下一轮训练的概率

样本的类不确定性越高,参与后续训练的概率越大,反之亦然. 基于式(1),样本参与训练的概率$p(\boldsymbol{x}_i)$

与$\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$为反比关系，反比关系的函数存在多样性，本文采用了带超参数的指数函数，如式(2)所示：

$$p(\boldsymbol{x}_i)\propto \mathrm{e}^{-\alpha\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|} \tag{2}$$

考虑到概率的约束，经过归一化后定义样本参与概率如下：

$$p(\boldsymbol{x}_i)=\frac{\mathrm{e}^{-\alpha\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|}-\mathrm{e}^{-\alpha}}{1-\mathrm{e}^{-\alpha}} \tag{3}$$

其中$p(\boldsymbol{x}_i)\in[0,1]$，当样本为简单样本或低质量样本时，$\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$较大，样本参与训练的概率$p(\boldsymbol{x}_i)$较小；当样本为边界样本时，$S_{\mathrm{t}}(\boldsymbol{x}_i)$与$S_{\mathrm{c}}(\boldsymbol{x}_i)$相对接近，$\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$较小，则样本参与训练的概率$p(\boldsymbol{x}_i)$较大.

超参数$\alpha$用于控制$p(\boldsymbol{x}_i)$对$\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$的敏感程度，从而控制参与训练的样本数量. 如式(4)和图3所示，不同$\alpha$概率函数形式不同. 当$\lim\limits_{\alpha\to-\infty}p(\boldsymbol{x}_i)=1$，会退化成常规的均权训练，所有样本都将加入训练. 当$\lim\limits_{\alpha\to 0}p(\boldsymbol{x}_i)=1-\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$，参与训练概率与$\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|$变为线性反比. 当$\lim\limits_{\alpha\to+\infty}p(\boldsymbol{x}_i)=0$，所有样本都不会被选中，提前结束训练. 当$\alpha$越小，被选中的样本越多.

$$p(\boldsymbol{x}_i)=\begin{cases}1, & \alpha=-\infty\\ 1-\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|, & \alpha=0\\ 0, & \alpha=+\infty\\ \dfrac{\mathrm{e}^{-\alpha\left|S_{\mathrm{t}}(\boldsymbol{x}_i)-S_{\mathrm{c}}(\boldsymbol{x}_i)\right|}-\mathrm{e}^{-\alpha}}{1-\mathrm{e}^{-\alpha}}, & \text{其他}\end{cases} \tag{4}$$

### 2.5 抽样训练的开启条件

当模型获得一定判决能力后才使用样本抽样策略，因此需要定义合适的启动条件. 一般情况下，训练过程的验证集损失随着训练的进行逐渐减少，相对变化量也逐渐变小，损失值达最佳收敛点后开始增大，产生过拟合. 因此，可通过判断每1轮结束后验证集收敛情况来衡量此时模型的性能，当验证集损失相对变化量小于阈值$\beta$时，样本抽样开启：

$$\frac{\left|\mathrm{pre}_{\mathrm{validLoss}}-\mathrm{cur}_{\mathrm{validLoss}}\right|}{\mathrm{pre}_{\mathrm{validLoss}}}<\beta \tag{5}$$

其中，$\mathrm{pre}_{\mathrm{validLoss}}$表示前1轮的验证集损失值，$\mathrm{cur}_{\mathrm{validLoss}}$表示当前轮的验证集损失值.

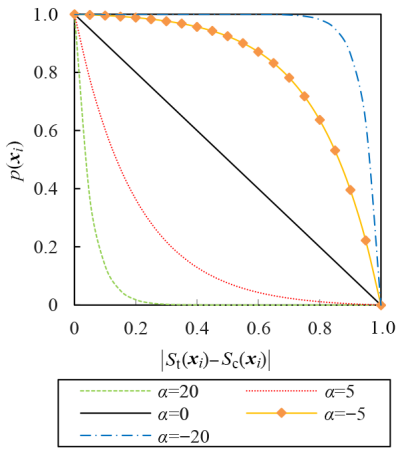



图3 不同$\alpha$时的参与训练概率

## 3 实验设计及结果分析

实验的主体目标是量化验证本文提出的CUS样本应用策略加速深度网络训练.

### 3.1 实验数据

实验使用AISHELL-1[24]和AISHELL-2[25]2个公开中文普通话语料库. 其中AISHELL-1包含178 h朗读式普通话数据，AISHELL-2总时长有1 000 h，内容涉及经济金融、智能家居和工业生产等多个领域. 在AISHELL-1小规模实验中，选取8个关键词；在AISHELL-2大规模实验中，选取70个2~3个汉字构成关键词，实验数据根据选取的关键词从上述2个数据集中抽取并划分，详情见表1.

表1 普通话数据集统计

| 数据集 | | 关键词样本数 | 关键词时长/h | 非关键词样本数 | 非关键词时长/h | 总样本数 | 总时长/h |
|---|---|---|---|---|---|---|---|
| AISHELL-1 | 训练集 | 18 022 | 24.4 | 54 066 | 67.3 | 72 088 | 91.7 |
| | 验证集 | 1 600 | 2.2 | 3 200 | 3.9 | 4 800 | 6.1 |
| | 测试集 | 3 200 | 4.4 | 6 400 | 7.9 | 9 600 | 12.3 |
| AISHELL-2 | 训练集 | 205 615 | 246.6 | 616 853 | 572.6 | 822 468 | 819.2 |
| | 验证集 | 15 167 | 17.3 | 30 334 | 28.1 | 45 501 | 45.4 |
| | 测试集 | 22 769 | 26.0 | 45 537 | 42.2 | 68 306 | 68.2 |

### 3.2 实验设置

语音关键词检测采用基于CTC(Connectionist Temporal Classification)的端到端CRNN-CTC方法，参数配置与文献[14]相同. 语音特征为80维的对数梅尔谱图，帧长25 ms，帧移10 ms. 模型结构如图4所示，其中CNN模块用于捕获局部信息，包含3个二维CNN(Con-

volutional Neural Networks)层,每层的卷积核大小均为3×3,步幅均为1,通道数分别为:16、32、32,并且每层之后添加了批标准化(Batch Normalization,BN)和ReLU(Rectified Linear Unit)激活函数,同时在第1个和第3个卷积层后使用最大池化,池化的大小为2,步幅为2. RNN模块用于学习时序信息,包含2层双向门控循环单元(Bidirectional Gated Recurrent Unit,Bi-GRU),均使用256个隐单元. 全连接模块包括2个全连接层,第1个全连接层的输入维度为256,输出维度为128,最后1个全连接层的输出维度与对应语音关键词的目标标签类别数一致.

训练阶段,在2个全连接层之间添加了比例为0.5的Dropout,使用初始学习率为0.001的Adam优化器,学习率每5 000步衰减1次,衰减系数为0.9,批大小为64. 验证集每轮评估1次,使用Early stopping策略,当模型在验证集上连续10轮没有提升时,停止训练,取其中性能最佳的模型作为最终模型. 实验平台使用Intel® Core™i7-6850K@3.6 GHz CPU与NIVIDIA GTX1080 TI GPU. 为保证实验数据的可靠性,对比实验在机器保证基本一致的环境下进行.

评价指标错误拒绝率(False Reject Rate,FRR)、虚警率越小表明性能越好,最大关键词加权值(Maximum Term Weighted Value,MTWV)[26]越大表明性能越好.

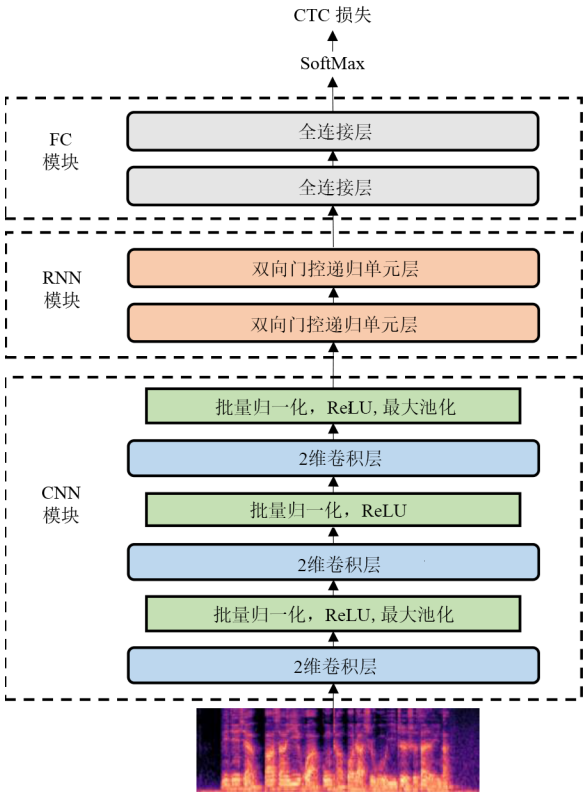



图4　CRNN-CTC方案

### 3.3　超参数优化

本文方法涉及2个重要的超参数:敏感参数$\alpha$和采样开启阈值$\beta$,其选值将影响CUS算法的效果. 本小节采取控制变量法,在AISHELL-1数据集上进行8个关键词的小规模实验完成2个超参数调优. 关键词为:中国(zhōngguó)、市场(shìchǎng)、公司(gōngsī)、百分(bǎifēn)、记者(jìzhě)、城市(chéngshì)、企业(qǐyè)和北京(běijīng). 在采用CUS训练过程中,评价指标采用FAR为0.5 FP/h时的测试集FRR,模型训练总耗时,以及模型收敛耗时(不包含收敛判断的10轮训练时间). 实验结果为5次不同随机种子进行训练的平均值.

敏感参数$\alpha$取值在$(-\infty,+\infty)$之间. 首先固定$\beta=0.1$,即验证集损失相对变化率小于10%时开启CUS,搜索优化$\alpha$参数,取值分别为:−100、−10、−5、−1、0、1、5、10、100. 表2给出不同$\alpha$值下模型性能与训练耗时.

当$\alpha=-\infty$,等同于一般常规训练方法;当$\alpha$从−100~100的变化过程中,训练时间逐渐缩短,FRR先下降后增长. $\alpha$越小,样本参与训练概率分布越接近1,每轮选中的样本越多,训练时间相对越长,但此时缺少对边界样本的关注度,性能只有微弱提升. 随着$\alpha$增大,被选中样本变少,训练时间变短,着重关注边界样本,进一步提升性能. 而当$\alpha$过大时,尽管此时训练时间进一步缩短,但被选中样本过少,性能逐渐变差. 当$\alpha=+\infty$,意味在开启采样之后,不选择任何样本,属于提前停止训练,训练时间最短,但模型未收敛,性能最差. 因此,需要权衡训练耗时和性能,选取合适的$\alpha$值. 在本数据集和训练条件下,当$\alpha=0$时,性能最佳,FRR为3.65%,训练耗时平均值为2 128 s,收敛耗时平均值为1 948 s.

对比训练耗时和收敛耗时可以看出,在开启CUS后二者的差距较小,在训练后期,参与训练的样本数量较少,最后10轮的训练耗时在训练总耗时中占比较小;不使用CUS,最后10轮训练耗时大,因此,训练耗时与收敛耗时差距较大. 本文方法的主要目的是降低深度网络模型的收敛时间,后续实验以“收敛耗时”作为主要比较指标.

抽样开启阈值$\beta$决定着CUS机制启动时间,经过$\alpha$参数调优,固定$\alpha=0$,对$\beta$参数搜索. $\beta$定义为验证集损失值相对变化率,因此范围应该在0~1,搜索取值分别为:0.05、0.1、0.25、0.5. 表3为不同$\beta$值下对应的实验结果. 当$\beta$设置较小时,开启的时间较晚,此时模型训练已接近收敛点,抽样机制不能充分发挥提高训练效率的作用,训练时间增长,性能收益下降. 当$\beta$较大时,过早进入抽样训练阶段,训练时间缩短,但此时模型的决策边界面不稳定,不能给样本进行可靠的类不确定性

度量，部分样本未能参与训练，最终性能变差.

综上所述，当敏感参数$\alpha=0$，采样开启阈值$\beta=0.1$时，能取得最佳的性能效果，具有较高的学习效率，后续实验以其作为CUS的超参数设置.

### 3.4 不同样本应用策略对比

比较对象包括常规训练策略和基于最大损失采样的选择性反向传播策略(Selective Backprop，SB)[21]，SB算法中的选择性参数设置为33%. SB算法[21]通过样本的损失值来决定是否进行反向传播计算，是固定训练样本集下与本文最接近的工作. 实验在小规模数据集上进行，模型方面除了使用CRNN-CTC方案，还使用了RNN-CTC方案[13]，验证CUS的泛化性. 表4给出CRNN-CTC和RNN-CTC 2种模型使用3种样本应用策略进行训练的结果，

**表2 在FAR为0.5 FP/h时，不同$\alpha$参数对应的实验结果($\beta=0.1$)**

| $\alpha$ | $-\infty$ | −100 | −10 | −5 | −1 | 0 | 1 | 5 | 10 | 100 | $+\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 训练耗时/s | 5 323 | 3 829 | 2 495 | 2 387 | 2 153 | 2 128 | 1 959 | 1 852 | 1 627 | 1 699 | 1 122 |
| 收敛耗时/s | 3 259 | 3 498 | 2 140 | 2 138 | 1 959 | 1 948 | 1 795 | 1 741 | 1 524 | 1 584 | — |

**表3 在FAR为0.5 FP/h时，不同$\beta$对应的实验结果($\alpha=0$)**

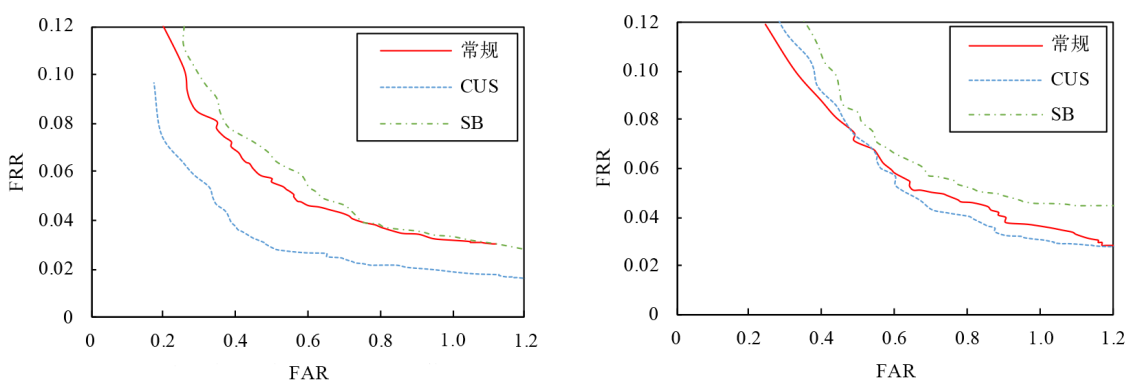
| $\beta$ | 0.05 | 0.1 | 0.25 | 0.5 |
|---|---|---|---|---|
| 训练耗时/s | 2 267 | 2 128 | 1 935 | 1 792 |
| 收敛耗时/s | 2 099 | 1 948 | 1 719 | 1 560 |
| FRR/% | 4.22 | 3.65 | 4.58 | 4.79 |

FRR同样是固定FAR为0.5 FP/h时取得. 图5($a$)和图5($b$)分别对应的检测误差权衡曲线(Detection Error Trade-off，DET).

在CRNN-CTC方案中，相对于常规训练策略，SB算法降低了训练时间，平均训练时长降至4 234 s，相对下降20.5%，平均收敛时长降至2 706 s，相对下降17.0%，性能方面却出现了明显的损伤，FRR升至5.79%，相对上升21.9%. CUS算法同样显著缩短了训练时间，平均训练时长缩短60.0%，平均收敛时长缩短47.5%，性能方面也有明显提升，FRR降至3.65%，下降30.1%，MTWV从0.837 4~0.853 1. 与文献[14]结论一致，RNN-CTC相对CRNN-CTC，由于没有使用CNN模块进行局部特征捕获，在3种样本应用策略下，其性能相对较低，且训练时间也相对较长，但3种样本应用策略的相对结果与CRNN-CTC一致，说明本文所提出的样本使用策略的效果与模型无关.

SB算法通过采样损失值大的样本进行梯度反向传播，提升训练速度. 然而，SB算法在整个训练过程中都使用最大损失值采样，训练开始时便使用最大损失抽样策略，尽管损失值大的样本能朝梯度更陡峭的方向下降，但此时模型并不具备较好的决策能力，损失值小的样本并不意味着其没有学习的意义，导致部分样本在整个训练过程中均没有参与，模型性能出现下降. 另外，损失值并不直接与语音关键词检测目标一致，损失值大的样本也不直接代表其在解码时容易出现漏检或虚警. 本文提出的CUS算法使用2阶段训练方式，通过类不确定性度量实现对易出现漏检或虚警的训练边界样本的直接筛选. 相对于常规训练策略，在基于CTC的端到端语音关键词检测方法中，使用CRNN-CTC和RNN-CTC不同方案时，CUS算法均能显著提高模型的训练效率，缩短训练时间的同时，提升检测能力.

表5为CRNN-CTC使用CUS算法时参与训练样本数的变化情况，表明CUS可使训练样本大大减少，降低了总体训练耗时.




($a$) CRNN-CTC实验结果 ($b$) RNN-CTC实验结果

图5 不同样本应用策略的DET曲线图

表4　不同样本应用策略的实验结果

| 方案 | 策略 | FRR/% | MTWV | 收敛耗时/s | 训练耗时/s |
|---|---|---|---|---|---|
| CRNN-CTC | 常规 | 4.75 | 0.837 4 | 3 259 | 5 323 |
| | SB | 5.79 | 0.823 7 | 2 706 | 4 234 |
| | CUS | 3.65 | 0.853 1 | 1 948 | 2 128 |
| RNN-CTC | 常规 | 7.59 | 0.794 0 | 7 670 | 10 477 |
| | SB | 8.3 | 0.745 7 | 5 819 | 7 778 |
| | CUS | 7.19 | 0.801 9 | 4 119 | 4 598 |

表5　开启采样后参与训练样本数变化情况

| 轮次 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| 数量 | 72 088 | 9 934 | 8 116 | 7 143 | 6 504 | 5 799 | 5 437 |

图6则为CRNN-CTC使用CUS算法时，利用t-SNE[27]数据降维方式，对开启采样后第1个轮次模型的决策可视化. 从图中可以观察到，此模型已具备一定决策能力，在不同关键词之间具有较明显的决策边界面. 通过利用类不确定性度量对各个样本参与训练概率的估计，大于0.85的高概率样本(红色圆点)主要集中在各个关键词的决策边界，而位于分布中心的样本参与训练概率相对较低. 表明CUS算法对位于决策边界的样本赋予了更高的参与训练概率，实现对边界样本的关注，在缩短训练时长的同时提升关键词检测能力.

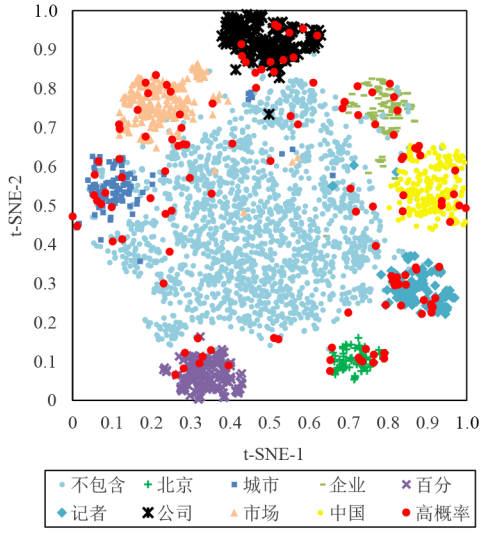



图6　模型的决策可视化

### 3.5　抗干扰特性

AISHELL-1具有足够高的可靠性[24]，字标注准确率95%以上，难以做到100%正确，因此，关键词训练样本的标注错误应是1个常见现象. 标注错误的样本参与训练是有害的. 在AISHELL-1数据集的基础上，构建1个带有错误标注的数据集，用以检验CUS具有降低错误标注样本损害的特点. 引入的错误标注可分为3类：

(1)不含关键词样本被标注为含关键词样本；

(2)含关键词样本被标注为不含关键词样本；

(3)含关键词样本之间标注错误. 错误标注样本集的构建做法是：随机抽取10%的含关键词样本(1 802个)和等量不含关键词的样本(1 802个)进行标注互换；随机抽取10%含关键词样本(1 802个)进行两两标注互换.

在新构建的带有错误标注的训练集中，含关键词样本中有3 604个属于错误标注，不含关键词样本中有1 802个属于错误标注，总共有5 406个错误标注，验证集和测试集保持不变. 实验采用CRNN-CTC模型，训练配置与模型参数设置与上文保持一致.

表6给出不同训练集配置下，3种样本应用策略的实验结果. 实验在3种不同配置训练集的情况下进行训练，其中，“完整”是指使用完整训练集(72 088个)进行训练；“剔除”是指在剔除错误标注后剩下训练集(66 682个)进行训练的结果；“包含”是指包含错误标注的训练集(72 088个)的实验结果.

对于常规训练策略，当训练集样本中错误标注样本被剔除，训练样本减少，与完整训练样本集相比性能降低，FRR升至6.61%，MTWV也降至0.814 6，训练耗时降低. 当训练集样本中包含错误标注时，错误标注将会带来明显性能损伤，FRR升至11.25%，MTWV降至0.747 3. 并且错误标注还会使模型在训练时难以收敛，由于早停机制的存在，训练提前结束，导致训练时间减少.

表6　不同样本应用策略的抗干扰实验结果

| 策略 | FRR/% | MTWV | 训练耗时/s | 备注 |
|---|---|---|---|---|
| 常规 | 11.25 | 0.747 3 | 4 272 | 包含 |
| | 6.61 | 0.814 6 | 3 401 | 剔除 |
| | 4.75 | 0.837 4 | 5 323 | 完整 |
| SB | 15.93 | 0.713 2 | 3 221 | 包含 |
| | 7.37 | 0.799 8 | 2 644 | 剔除 |
| | 5.79 | 0.823 7 | 4 234 | 完整 |
| CUS | 7.09 | 0.794 3 | 1 899 | 包含 |
| | 4.43 | 0.855 2 | 1 725 | 剔除 |
| | 3.65 | 0.853 1 | 2 128 | 完整 |

对于SB算法，训练样本的减少同样使模型检测性能下降，FRR升至7.37%，对应的MTWV降至0.799 8，训练耗时也相应缩短. 在使用包含错误标注的训练集时，由于SB算法的核心是关注损失值更高的样本，错误标注样本在训练过程中会产生较高损失值，使模型更多地对错误标注样本进行学习，最终导致模型出现严重的性能损伤，FRR显著升至15.93%，MTWV降至0.713 2，同样由于模型难以收敛，早停机制使训练时间

减少.

对于CUS算法,在剔除错误标注数据样本后性能下降,FRR升至4.43%,MTWV升至0.855 2,训练耗时降低.在包含错误标注时性能也进一步下降,但CUS算法能减轻错误标注带来的性能损伤,FRR仅升至7.09%,对应的MTWV为0.794 3,训练时长略微降低.

表7为使用CUS进行训练时,3种类型的错误标注样本在开启抽样训练后,参与训练的概率大于0.5的个数变化情况,第6轮时开启了类不确定性抽样训练,模型最佳性能的收敛点在第11轮.

**表7 错误标注样本中参与训练的概率大于0.5的数量**

| 轮次 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|
| 第1类错误 | 1802 | 75 | 54 | 41 | 40 | 39 | 42 |
| 第2类错误 | 1802 | 775 | 604 | 531 | 526 | 513 | 563 |
| 第3类错误 | 1802 | 643 | 487 | 459 | 471 | 477 | 499 |

从统计结果可知,在开启CUS抽样训练之后,错误标注样本的参与训练概率大于0.5的个数明显少于1 802个.显然,参与抽样训练阶段的错误标注样本量明显变少,一定程度上缓解了错误标注样本带来的损伤.

在本训练集中,不含关键词的样本较多,因此,模型前期能明确判决不含关键词的样本.对于第1种错误类型,由于将不含关键词样本标注成含关键词样本,对于含关键词这一标注而言,目标类得分非常低,而竞争类得分非常高,即类不确定性低,这类样本在采样阶段基本都能被筛除,在训练第6轮时参与训练概率大于0.5的个数不到80个.如前所述,CUS区分错误标注样本前提是模型已具备一定决策能力,对于第2种和第3种错误类型,两者的正确标注都是含关键词的样本.然而训练集中含关键词的样本相对少,加上错误标注样本带来一定的性能损伤,模型对含关键词样本的决策能力相对较差.导致在开启采样时存在部分难以判决的含关键词样本,使部分错误标注样本类不确定性得分较高.所以在采样阶段多数含关键词的错误标注样本被筛除,但还是有一定数量的样本会继续进入训练.从变化趋势和最终收敛情况可以看出,进入采样阶段后,模型的决策能力会随着训练的加深进一步加强,参与训练概率较高的错误标注样本数量变少.达到最佳收敛点之后,由于模型继续拟合那一部分未能被筛除的错误标注样本,导致模型决策产生一定误判,决策能力下降,这时高概率的错误标注样本数量将会逐渐增多,偏离最佳收敛点.

综上所述,当训练集中存在一定数量的错误标注样本时,CUS能利用已建立的良好决策面,利用类不确定性抽样筛除错误标注样本,缓解错误标注样本带来的性能损伤.这表明CUS具有屏蔽部分标注错误样本干扰的能力,提高低质量数据集利用率和模型的训练效率.

### 3.6 在大规模数据集上的表现

选取AISHELL-2数据集上70个关键词进行实验,在大规模数据集及多关键词的条件下对CUS算法进行验证.在CRNN-CTC上进行实验,训练配置和其余模型参数设置与上文保持一致,仅改变最后1层全连接层的输出维度.

表8为3种算法在大规模数据集上的实验结果.在训练时长方面,由于数据量增多,且多个关键词使模型输出单元增多,常规训练策略的收敛速度慢,平均训练时间达到55.9 h.采用SB算法后,训练时长降为40.8 h,相对减少27.0%.性能上,当FAR为0.5 FP/h时,FRR有较小幅度降低,由3.69%降为3.58%,相对下降2.98%.但MTWV变差,由0.912 8降至0.902 7.而CUS显著降低了训练耗时,平均训练时长仅需30.7 h,相对常规方法减少45.1%.性能方面也有改善,当FAR为0.5 FP/h时,FRR由3.69%下降至3.24%,相对下降12.2%.MTWV由0.912 8升至0.916 4.由于数据规模增大,模型学习到足够的信息进行决策,此时CUS算法与常规训练策略相比性能收益差距缩小.实验表明在大规模数据场景中,CUS能显著缩短关键词检测模型的训练耗时,同时取得略微性能提升.

**表8 大规模数据集的实验结果**

| 策略 | FRR/% | MTWV | 训练耗时/h |
|---|---|---|---|
| 常规 | 3.69 | 0.912 8 | 55.9 |
| SB | 3.58 | 0.902 7 | 40.8 |
| CUS | 3.24 | 0.916 4 | 30.7 |

## 4 总结

在给定训练样本集时,本文针对端到端语音关键词检测模型训练效率问题,提出基于样本类不确定性抽样(CUS)的样本应用策略,核心是在模型训练中后期对语音关键词样本进行类不确定性度量,动态调整样本参与下1轮训练的概率,相对加大边界样本的训练力度,对易造成漏检或虚警的样本给予更多关注,减少简单样本的重复学习和屏蔽低质量样本的干扰,实现训练样本的高效应用,提高模型训练效率.实验表明CUS算法有效加速端到端语音关键词检测模型的训练进程,显著提高模型训练效率,获得好的检测性能.低质量训练样本实验表明,该方法能屏蔽掉大部分错误标注样本,提高模型的泛化能力.

本文方法可认为是CL方法基本思想的一种实现方案,在性能和收敛速度两方面均实现了CL的期望效果.另外本文方法具有即插即用特点,与深度网络结

构、训练方法没有直接关系,因此具有广泛的应用前景.

**参考文献**


[1] LIU W C, HUANG Z H, WANG D F. Keyword spotting based on efficient neural architecture search[C]//2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML). Piscataway: IEEE, 2023: 432-436.

[2] KUROKAWA T, KAI A. Robust query-by-example spoken term detection for unknown words using speech retrieval-oriented E2E ASR modeling[C]//2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). Piscataway: IEEE, 2021: 316-317.

[3] NA Y Y, WANG Z T, WANG L, et al. Joint ego-noise suppression and keyword spotting on sweeping robots[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 7547-7551.

[4] LI M R. A lightweight architecture for query-by-example keyword spotting on low-power IoT devices[J]. IEEE Transactions on Consumer Electronics, 2023, 69(1): 65-75.

[5] 田颖慧, 贺前华, 郑若伟, 等. 基于特征空间轨迹信息的语音关键词检测方法[J]. 电子学报, 2023, 51(10): 2915-2924.
TIAN Y H, HE Q H, ZHENG R W,et al. Spoken term detection based on feature space trajectory information[J]. Acta Electronica Sinica, 2023, 51(10): 2915-2924. (in Chinese)

[6] CHEN G G, PARADA C, HEIGOLD G. Small-footprint keyword spotting using deep neural networks[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2014: 4087-4091.

[7] TIAN Y, YAO H T, CAI M, et al. Improving RNN transducer modeling for small-footprint keyword spotting[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 5624-5628.

[8] PETER D, ROTH W, PERNKOPF F. End-to-end keyword spotting using neural architecture search and quantization[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 3423-3427.

[9] CHOROWSKI J K, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]//28th International Conference on Neural Information Processing Systems (NIPS). Montreal: MIT Press, 2015: 577-585.

[10] 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-ICML. New York: ACM, 2006: 1143844.

[11] SHAN C H, ZHANG J B, WANG Y J, et al. Attention-based end-to-end models for small-footprint keyword spotting[C]//Interspeech 2018. Baixas: ISCA, 2018: 2037-2041.

[12] HIGUCHIL T, GUPTA A, DHIR C. Multi-task learning with cross attention for keyword spotting[C]//2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE, 2021: 571-578.

[13] BAI Y, YI J Y, NI H, et al. End-to-end keywords spotting based on connectionist temporal classification for Mandarin[C]//2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). Piscataway: IEEE, 2016: 1-5.

[14] YAN H K, HE Q H, XIE W. CRNN-CTC based mandarin keywords spotting[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 7489-7493.

[15] LAN X T, HE Q H, YAN H K, et al. A novel re-weighted CTC loss for data imbalance in speech keyword spotting[J]. Chinese Journal of Electronics, 2023, 32(3): 465-473.

[16] SHIN D, KIM G, JO J, et al. Low complexity gradient computation techniques to accelerate deep neural network training[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 5745-5759.

[17] NIKOLOV M, TSENOV G, NAKOV O, et al. Application of GPU accelerated deep learning neural networks for COVID-19 recognition from X-ray scans[C]//2022 10th International Scientific Conference on Computer Science (COMSCI). Piscataway: IEEE, 2022: 1-5.

[18] KIRTAS M, PASSALIS N, TEFAS A. Multiplicative update rules for accelerating deep learning training and increasing robustness[J]. Neurocomputing, 2024, 576: 127352.

[19] ZHANG C, ÖZTIRELI C, MANDT S, et al. Active mini-batch sampling using repulsive point processes[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2019, 33(1): 5741-5748.

[20] PENG X Y, LI L, WANG F Y. Accelerating minibatch stochastic gradient descent using typicality sampling[J].

IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(11): 4649-4659.
[21] JIANG A H, WONG D L K, ZHOU G, et al. Accelerating deep learning by focusing on the biggest losers[EB/OL]. (2019-10-02)[2024-01-10]. http://arxiv.org/abs/1910.00762.
[22] CAO R Y. Towards accelerated and robust rreinforcement learning with transfer learning[C]//2022 International Conference on Big Data, Information and Computer Network (BDICN). Piscataway: IEEE, 2022: 335-340.
[23] WANG X, CHEN Y D, ZHU W W. A survey on curriculum learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4555-4576.
[24] BU H, DU J Y, NA X Y, et al. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment. Piscataway: IEEE, 2017: 1-5.
[25] DU J Y, NA X Y, LIU X C, et al. AISHELL-2: Transforming mandarin ASR research into industrial scale[EB/OL]. (2018-09-13)[2024-01-10]. http://arxiv.org/abs/1808.10583.
[26] WEGMANN S, FARIA A, JANIN A, et al. The TAO of ATWV: Probing the mysteries of keyword search performance[C]//2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway: IEEE, 2013: 192-197.
[27] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

## 作者简介

**贺前华** 男，1965 年 2 月出生，湖南邵东人. 华南理工大学教授、博士生导师. 主要研究方向为智能音频信号处理、语音识别和说话人识别.
E-mail: eeqhhe@scut.edu.cn

**陈永强** 男，1999 年 3 月出生，四川巴中人. 华南理工大学硕士研究生. 主要研究方向为语音关键词检测.
E-mail: 202221012416@mail.scut.edu.cn

**郑若伟** 男，1998 年 2 月出生，广东汕头人. 华南理工大学硕士研究生. 主要研究方向为语音关键词检测、语音识别.
E-mail: ruoweizheng@foxmail.com

**黄金鑫** 男，2001 年 6 月出生，广东河源人. 华南理工大学硕士研究生. 主要研究方向为语音关键词检测、声音事件检测.
E-mail: eejinxinhuang@mail.scut.edu.cn