关键词识别综述

林正青[1)] 赵泳豪[2)] 罗天辰[2)]

[1)](南开大学 网络空间安全学院, 天津市 300350)
[2)](南开大学 计算机学院, 天津市 300350)
[)](南开大学 计算机学院, 天津市 300350)

摘 要 关键词识别（Keyword Spotting，KWS）是一类最简单的语音识别任务，主要应用于低资源设备上的设备唤醒场景。经典解决方案是使用隐马尔可夫模型（HMM），现在常用的方法是构建和测试轻量级的端到端神经网络模型，其中以卷积网络的应用为首。本文首先介绍 KWS 任务中应用较多的训练数据集，接下来依次介绍 KWS 系统所使用的经典统计方法 HMM 和端到端方法常用的神经网络结构，并提及某些有效的训练策略，最后阐述可能的未来方向。

关键词 关键词识别；卷积神经网络；谷歌语音识别数据集；循环神经网络；注意力机制
中图法分类号 TP391 DOI 号 10.11897/SP.J.1016.01.2022.00001

A Survey of Keyword Spotting

LIN Zhengqing[1)] ZHAO Yonghao[2)] LUO Tianchen[2)]

[1)](Department of Cyber Security, Nankai University, Tianjin 300350, China)
[2)](Department of Computer Science, Nankai University, Tianjin 300350)
[)](Department of Computer Science, Nankai University, Tianjin 300350)

Abstract Keyword Spotting (KWS) is one of the simplest speech recognition tasks, and is mainly used in device wake-up scenarios on low-resource devices. The classic solution to this task is to use Hidden Markov Models (HMMs), and now the common approach is to build and test lightweight end-to-end neural network models, led by the application of convolutional networks. This paper first introduces the training datasets that are widely used in KWS tasks, then introduces the classical statistical method HMM used by the KWS system and the neural network structure commonly used in the end-to-end method, and mentions some effective training strategies, and finally elaborates possible.

Key words Keyword Spotting; CNN; Google Speech Command Dataset; RNN; Attention

1 引言

1.1 任务描述

关键词识别（Keyword Spotting，KWS）是一类简单的语音识别任务，主要目标是构建和测试小型的模型，使之可以从一组十个或更少的目标词中检测何时说出一个词，同时尽可能少地从背景噪声或不相关的语音中误报。

大多数语音接口在本地设备上运行识别模块，持续监听来自麦克风的音频输入，而不是通过互联网将数据发送到云端，而一旦听到触发信号，就会开始将音频传输给 Web 服务。在本地识别触发关键词的设计是考虑到初始识别通过网络发送音频数据的成本和隐私风险，以及服务器响应延迟可能造成的不佳用户体验。

由于本地设备的存储和总计算能力通常远低于大多数服务器，要想较为实时地进行交互式响应，设备上的模型应当轻量，并需要尽可能少的计算。对于电量有限的移动设备，这种持续运行的程序应非常节能，以避免用户感到设备电量耗尽太快；对于插入式电设备，使用轻量级模型也有助于降低功耗。

2 问题与挑战

从 KWS 的应用场景出发，可以总结出该任务面临的主要问题和挑战：

收稿日期：2022-11-23；修改日期：2022-11-23 林正青，女，本科学历，学生，非计算机学会 (CCF) 会员，主要研究领域为语音识别和自然语言处理. E-mail: . 赵泳豪，性别男，本科学历，学生，非计算机学会 (CCF) 会员，主要研究领域为语音识别和自然语言处理. E-mail: . 罗天辰，性别女，本科学历，学生，非计算机学会 (CCF) 会员，主要研究领域为语音识别和自然语言处理. E-mail: .

2.1 设计轻量级模型

实际 KWS 系统的一个重要特点是其尺寸受运行平台内存和处理能力的限制。参数过多的模型比如较深的 RNN、LSTM 等将不适用该任务，虽然它们也可以取得不错的效果。目前比较出色的方法可以将模型参数所占据的内存限制在 100K 以内。

2.2 存在干扰——噪音（noise）、远场和低录音质量

噪音通常指关键语音之外的背景声音，可以分为两类：背景噪声和不相关语音。噪音可造成误报，或在识别关键词时产生干扰。KWS 系统的大部分输入是静音或背景噪音，而不是语音，并且无关语音也不应当触发识别。KWS 系统往往在纯净语音环境下表现非常好，一旦存在噪音，系统性能就将显著下降。为提高对背景噪声的鲁棒性，近年来主要有两类解决方案，即多条件训练、前端增强和数据增强。多条件训练汇集不同环境下的数据训练神经网络以获得更健壮的系统，但以这种方式学习的特征表示以及 KWS 的性能仍比预期的要差，因为 KWS 网络的大小受到平台内存和处理能力的限制。前端增强技术在将噪声流中的干扰信号馈送到 KWS 系统之前使用滤波技术将其过滤掉。以上两者更多使用信号处理技术，而数据增强指给原始录音叠加背景噪声或其他语音 [4]，可以增强模型对噪音的鲁棒性。

2.3 未标记的声音起始点

起始点不明确会加大误报的可能性。实际场景下语音的起始点是没有标明的，而考虑到手动标注的成本，训练数据中的音频往往也不会标记出起始点，尽管可以在收集录音时加以处理使它们的开始点近似对齐。

2.4 可否支持用户自定义关键词

大多数采用语音助手的设备都使用预定义的唤醒词（关键字），例如“Hey Siri”、“OK Google”或“小雅小雅”。作为个性化设备的一种方式，能够任意定义关键字可能很有吸引力，但实现起来会有一定挑战性，原因是自定义的词语可能不在训练的分布范围内。此外如果是面向其他语言进行” 从零到一” 的工作，可能还需要构建对应语言的数据集。

3 研究现状分析

3.1 主要数据集

3.1.1 Google Speech Commands Dataset

这是最近几年 KWS 任务中使用最广泛的数据集，现分为 v0.01 和 v0.02 两个版本，语言均为英语。数据集 v1（v0.01）包含来自 1881 位录音提供者的 64727 条录音，共包括 105829 条长度 1 秒且声音起始点近似对齐的关键词语音，以 16 KHz 采样率编码为 16 位单通道 PCM 值。录音提供者共 2618 位，每个人都有唯一的随机生成的八位十六进制标识符。未压缩的数据集总文件在磁盘上占用大约 3.8 GB，可以存储为 2.7GB gzip 压缩的 tar 存档，内部已分为训练、验证和测试三个 txt 文件，每条录音在不同版本中总保持在同一集合中，由使用其名称生成的散列函数值决定。为未知语音标记 Unknown Word，静音标记 Silence。该数据集在 Creative Commons BY 4.0 开源许可下发布，因此可以出现在商业场景里，可以被非大型机构人士使用，也能够整合到其他脚本中。一些主流 ML 框架可以调用该数据集，比如 Pytorch 中 torchaudio.dataset 模块的 SPEECHCOMMANDS 类（默认使用 v2（v0.02），可更换）。

3.1.2 其他数据集

除了上述谷歌数据集是一个短录音片段集合，其他数据集最小都以句子为单位。在主流论文里出现频率相对高的有 Hey Snips（关键词为”Hey Snips”）等。使用规模较大、含有更多词汇的数据集进行训练，能够提升模型可拟合的分布范围，可以尝试用于用户自定义关键词功能的实现。

3.2 评价指标

测试模型预测精确程度时现有文献中最常用的指标依次是准确率（Accuracy）、错误拒绝率（False Rejected Rate）或 F1 值（F1 measure，是精确率（precision）和召回率（recall）的调和平均值），计算方式如下，其中 A、R 分别代表 Accepted 和 Rejected：

$$ACC = \frac{TA + TR}{NUM\_ALL\_SAMPLES}$$

$$FRR = \frac{FR}{NUM\_ALL\_TESTS}$$

$$F1 = \frac{2}{\frac{1}{precise} + \frac{1}{recall}} = \frac{2TA}{2TA + FA + FR}$$

而衡量 KWS 系统的实时性时可以使用实时率指标。

3.3 输入数据处理和表示

通常使用 MFCC 技术（一种依照人类听觉对数曲线进行滤波的技术）处理输入音频，得到梅尔频率倒谱图后以几十毫秒的区间分帧，每个帧可以作为一张待处理图像。

3.4 经典方法：GMM/DNN+HMM

被用于 KWS 任务并取得较好效果的一种经典统计方法是隐马尔可夫模型（Hiden Markov Model, HMM），而且直到现在这种方法还有应用。在使用经典方法的 KWS 系统里，HMM 被用来建模时序数据，而 GMM（Gaussian Mixture Model，高斯混合模型）提供各给定语音帧的先验状态概率。后来神经网络开始替代 GMM 的功能：传统 DNN 使用全连接层预测先验状态概率，而 HMM 解码器结合多个语音帧的 DNN 预测计算各状态链得分，以某种策略（比如 Beam Search）做选择，然后与关键词进行比对。这种方法的缺陷首先在于计算量较大，此外 HMM 的各状态之间独立的假设也未必成立，因为语音前后各帧一般是有语义联系的，这会影响系统性能。当端到端神经网络方法被提出后，由于其出色的预测效果、可降低的计算成本和简便性，最近的 KWS 系统更多采用这类方法。

3.5 端到端方法最常用的神经网络结构：CNN

CNN 模型参数较 RNN 类少很多（约 10^5 数量级），适用于需要轻量级模型的场合。在现今的 KWS 系统中 CNN 已被实际应用，并显示了出色的准确性。许多模型仍然使用普通的卷积层（以及池化层）作为基础网络架构，比如 2021 年提出的 Matchbox Net，参数量 93K，达到了大约 97.48% 的准确率。CV 领域提出的 VGG 也被用于 KWS 任务，并获得了较好的效果（根据 2018 年测试数据，VGG19+Batch Normalization 在谷歌数据集 v1 上获得 97.34% 的准确率）。卷积核可以提取局部特征，适于发掘很小局部之间的相关性，但由于其大小有限，不能照顾到较长期的空间关系和时序依赖，但从模型实际测试结果来看，对 KWS 中这些较短关键词的识别还是可以达到较好的效果。

3.6 时序卷积 (Temporal Convolution Network, TCN)

基于 CNN 结构，针对时序依赖性进行的改进。提出“因果卷积”（Causal Convolution）和“膨胀卷积”（dialated convolution）两类层间依赖模式，使得模型能够参考过去更长一段时间的信息，但感受野理论范围不如 Transformer 大。

3.7 CTC

联结主义时间分类（Connectionist Temporal Classification, CTC）是一种被用于解决输入输出对应问题的算法，基于条件独立（指各帧之间互不相关，之前介绍 HMM 时已提到）、输入输出单调对应和输入长度必须大于标签长度三个假设，性能并非最好，现在大多不被单独使用。一个使用思路是将其 loss 作为补充加入模型 loss。

3.8 ResNet

ResNet 结构中 identity mapping 的跨层连接结构可以解决梯度消失问题。但单纯 ResNet 网络每提升一点准确率需要叠加几乎一倍的层数，要达到最先进性能也需要 10^5 数量级的参数。事实上，现有方法中结合卷积层和几个 ResNet 连接时可以获得很好的效果。

此外，WRN（Wide ResNet）网络是针对 ResNet 层数过多问题的一个改进方案。WRN 减少网络深度增加宽度，使用约两倍参数就可以提升一半性能，在 2018 年的测试中，WRN-28-10 网络的准确率达到 97.94%，性能几乎是最优的。

3.9 RNN-Transducer（RNN-T）

这种结构的一个应用来自 [7]。这篇文章旨在构建支持任意关键字的系统，并指出仅是基于 RNN-T 的系统对于没出现在训练数据里的音素序列表现较差（不适应低资源环境），提出的解决方案是引入 TTS 技术为没有任何自然语言训练数据的关键字合成语音（即人工合成一段数据用于训练），采用一个预训练的 RNN-T 子模型，最终性能有相当大的改进。该文章还包括使用上述方法研究说话人多样性、噪声仿真和同一说话人不同数量的合成语音对于 KWS 性能的影响，结论是更多的说话者和每个说话者更多的数据有利于 KWS 性能，而不对 TTS 生成的数据执行噪声和房间脉冲响应模拟会对 KWS 性能产生负面影响。

3.10 基于 Attention

类似 Transformer 这样仅使用 self-attention 机制的模型可以捕捉长距离信息，但相应地会弱化局部特征信息。如今很多文献在采用卷积层 +Attention（可能是 multi-look 型）的结构。

3.11 Conformer

2020 年提出的结构，结合了卷积层和 Transformer，二者可以优势互补。该模型在语音识别任务上效果较好。

3.12 学习策略

本节特别指出 KWS 任务中一些常用的效果较好的学习策略。

3.12.1 批量规范化（Batch Normalization）

为保证隐藏层输入数据的质量，防止由于输入数据本身或带来的梯度过大导致神经元参数“死亡”（梯度消失），批量规范化方法对隐藏层的输出做归一化。在卷积层后面加上 BN 层的方式被很多文章所采用，目前效果较好的文章里也几乎都加入了 BN 层。

3.12.2 神经架构搜索（NAS）

自动化 KWS 任务的神经网络架构设计，已在图像分类和语言建模任务被应用和评估。可以参考[9]。

3.12.3 模型压缩

现有工作中已经有使用量化等模型压缩方法来寻找等效小规模模型的例子。量化指降低模型参数的存储位数，可以在保证性能目标的基础上加入此步骤压缩体量较大的模型。

3.13 关键数据不平衡

训练数据集中正样本远少于负样本。在这种数据分布上训练的模型在样本较少的类上表现不佳。目前最有效的解决方案是重新加权训练损失，也称为成本敏感学习（cost-sensitive learning），可参考[12]。

4 未来方向

考虑到用户自定义关键词的任务难度，这仍然是个值得探索的应用方向。近两年已有一些此方向的研究，使用

对声音频谱/倒谱图的处理可与图像处理相类比，一个直接的思路是可以继续根据 KWS 任务的特点引入后者领域里合适的前沿模型和算法，加以组合、调优和测试，产生新的解决方案。期间可以使用 NAS 和模型压缩技术，结合经验和数据推理结果进行轻量级网络的高效构建。

目前没有找到在 KWS 任务上应用 Conformer 的文献，在接下来的实验环节或许可以测试该模型性能。

5 结论

本文介绍了 KWS 任务及其主要应用场景，描述了在这些场景下解决方案面临的问题和挑战，然后介绍了任务中应用较多的训练数据集，依次简要解释 KWS 系统所使用的经典统计方法 HMM 和端到端方法常用的神经网络结构，并提及某些有效的训练策略。最后从新的应用场景（即用户自定义关键词）和如何设计新的解决方案两方面阐述了有探索意义的未来方向。

参考文献

[1] WARDEN P. Speech commands: A dataset for limited-vocabulary speech recognition[J]. arXiv preprint arXiv:1804.03209, 2018.

[2] RABINER L R. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77 (2):257-286.

[3] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

[4] MAJUMDAR S, GINSBURG B. Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition[J]. arXiv preprint arXiv:2004.08531, 2020.

[5] LI X, WEI X, QIN X. Small-footprint keyword spotting with multi-scale temporal convolution[J]. arXiv preprint arXiv:2010.09960, 2020.

[6] SHARMA E, YE G, WEI W, et al. Adaptation of rnn transducer with text-to-speech technology for keyword spotting[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2020: 7484-7488.

[7] LIU Z, LI T, ZHANG P. Rnn-t based open-vocabulary keyword spotting in mandarin with multi-level detection[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2021: 5649-5653.

[8] 俞栋, 邓力, 俞凯, 等. 解析深度学习: 语音识别实践.[M]. 北京: 电子工业出版社, 2016.

[9] MO T, YU Y, SALAMEH M, et al. Neural architecture search for keyword spotting[J]. arXiv preprint arXiv:2009.00165, 2020.

[10] BAE J, KIM D S. End-to-end speech command recognition with capsule network.[C]//Interspeech. [S.l.: s.n.], 2018: 776-780.

[11] HUANG J, GHARBIEH W, SHIM H S, et al. Query-by-example keyword spotting system using multi-head attention and soft-triple loss [C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2021: 6858-6862.

[12] HOU J, SHI Y, OSTENDORF M, et al. Mining effective negative training samples for keyword spotting[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2020: 7444-7448.

[13] YU M, JI X, WU B, et al. End-to-end multi-look keyword spotting[J]. arXiv preprint arXiv:2005.10386, 2020.

[14] JUNG M, JUNG Y, GOO J, et al. Multi-task network for noise-robust keyword spotting and speaker verification using ctc-based soft vad and global query attention[J]. arXiv preprint arXiv:2005.03867, 2020.

[15] RYBAKOV O, KONONENKO N, SUBRAHMANYA N, et al. Streaming keyword spotting on mobile devices[J]. arXiv preprint arXiv:2005.06720, 2020.

[16] LIU H, ABHYANKAR A, MISHCHENKO Y, et al. Metadata-aware end-to-end keyword spotting.[C]//INTERSPEECH. [S.l.: s.n.], 2020: 2282-2286.

[17] YILMAZ E, GEVREK O B, WU J, et al. Deep convolutional spiking neural networks for keyword spotting[C]//Proceedings of INTERSPEECH. [S.l.: s.n.], 2020: 2557-2561.

[18] YANG C, WEN X, SONG L. Multi-scale convolution for robust keyword spotting.[C]//INTERSPEECH. [S.l.: s.n.], 2020: 2577-2581.

[19] ZHANG P, ZHANG X. Deep template matching for small-footprint and configurable keyword spotting.[C]//INTERSPEECH. [S.l.: s.n.], 2020: 2572-2576.

LIN Zhengqing, Bachelor, Student, Her research interests include Speech Recognition and NLP.

ZHAO Yonghao, Bachelor, Student, His research interests include Speech Recognition and NLP.

LUO Tianchen, Bachelor. Student. Her research interests include Speech Recognition and NLP.