多实例级联分类在窄带音频中的关键词检测

Ahmad Abdulkader
Voicera
ahmada@voicera.ai

Kareem Nassar
Voicera
kareemn@voicera.ai

Mohamed Mahmoud
Voicera
geish@voicera.ai

Daniel Galvez
Voicera
dt.galvez@gmail.com

Chetan Patil
Voicera
chetanp@voicera.ai

Abstract

我们建议使用级联分类器来处理窄带（NB），8kHz 音频中的关键词检测（KWS）任务——该任务在非独立同分布环境下的挑战性大于大多数最先进的 KWS 系统所面对的任务。我们提出了一种结合深度神经网络（DNNs）、级联、多特征表示和多实例学习的模型。级联分类器处理了任务中类别不平衡的问题，并通过早期终止来减少计算受限设备上的功耗。该 KWS 系统在每小时假阳性率为 0.75 的情况下，实现了 6%的假阴性率。

1 介绍和问题描述

在 Voicera，我们正在构建 Eva——企业语音助手——以通过语音与会议参与者进行协作 [1]。我们认为与 Eva 的互动应该感觉像与其他任何会议参与者一样自然，因此我们设计了 Eva 来识别并响应语音命令。Eva 在会议期间持续聆听对话，并口头确认唤醒词 “Okay Eva”。为了使这种互动感觉自然，Eva 的关键词检测（KWS）和可听见的回应需要是实时的。

Eva 持续预测实时音频流中是否说出感兴趣的关键字。为了避免用户因不请自来的中断而觉得 Eva 烦人，假阳性率应低于每小时一次；另一方面，如果提到 Eva 时它没有回应，用户可能会放弃使用该服务，因此我们在保持低假阳性率的同时力求最大化召回率。

除了其他关键词搜索系统在处理实时语音时面临的挑战外，Eva 的关键词搜索系统——为了实现无处不在的目标——需要支持通过公共交换电话网络传输的语音信号，这通常使用 G.711：一种以低比特率和 8kHz 采样率运行的窄带音频编解码器 [2]。当听到窄带音频时，无论是人类还是自动语音识别都会显著地丧失准确性 [2, 3, 4]。此外，Eva 的关键词搜索系统需要适应新麦克风、环境设置、说话人和噪声配置文件的特点——这些特点差异巨大——使得输入到关键词搜索系统的信号是非独立同分布的。

2 现有技术

语音启用的 AI 助手——如苹果的 Siri、亚马逊的 Alexa 和谷歌助手——执行类似的关键词搜索任务，以使用户能够与其设备进行交互。谷歌提出了一种使用卷积神经网络来执行关键词

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

搜索任务以检测 14 个不同短语 [5]; 该架构与以前使用深度神经网络 (DNN) 的方法相比，误报率相对降低了 27-44% [6]，而后者相对于使用隐马尔可夫模型 (HMM) 的关键词搜索方法，在置信度得分上提高了 45% [7]。在 [8] 中，苹果公司使用一个 DNN 每 10 毫秒预测 20 个声音类别的分数，并将这些分数结合在一个类似 HMM 的图中计算综合分值来指示是否说出“嘿 Siri”。与我们在下文详细介绍的方法相似，作者使用了两个分类器的级联以节省电力。鉴于苹果、亚马逊和谷歌对其硬件规格的控制能力，他们能够获取宽带音频，这显著提高了语音识别系统的准确性。

3 方法

我们对 KWS 问题的方法如图 1 所示。KWS 系统每 10 毫秒触发一次。捕捉过去 500 毫秒的音频信号进行处理，因为我们发现命令的中位持续时间是 500 毫秒。在本文中，我们通常将这 500 毫秒窗口称为“示例”。我们的六个分类器——由两个基于不同特征表示训练的 DNN 组成的三个级联分类器——都是具有两个 128 输入隐藏层的全连接 DNN。

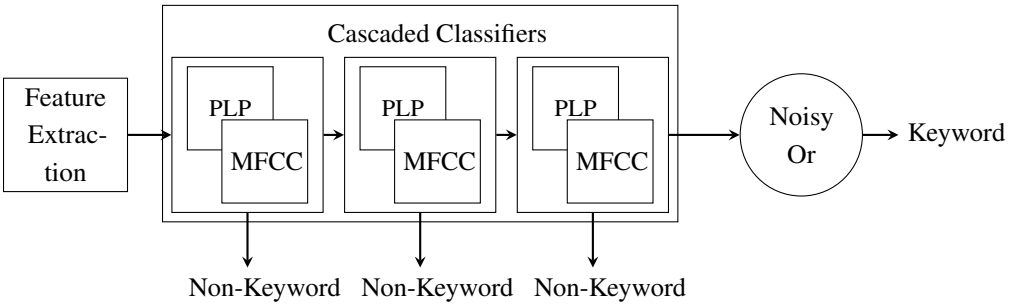

图 1: 关键单词系统示意图：(i) 特征提取 (ii) 级联分类器 (iii) 噪声或

3.1 数据

我们最初从 200 名个人中通过众包平台收集了 19k 个用户示例。这些示例涵盖了居住在美国大陆的各种说话人。这些示例在我们的实验中充当正例。每个级联分类器的反例都是从不包含关键词的各种会议录音音频样本库中生成的。所有的音频示例要么作为窄带，8kHz 音频获取，要么转换为窄带，8kHz 音频。

80%的数据被指定用于训练；剩下的 20%用于评估。数据根据个别说话人进行了分层：一个说话人要么属于训练集，要么属于评估集。

3.2 多表示特征提取

充分采样负样本具有挑战性。因此，判别分类器会遭受所谓的“新颖检测”问题：在训练过程中未遇到的声音可能会被误分类为关键词——导致假阳性。这是因为由判别分类器学习到的决策边界在训练过程中未采样的区域是未定义的。

为克服这一点，我们在两种不同的表示方法上训练了两个分类器，分别是梅尔频率倒谱系数（MFCC）和感知线性预测（PLP）特征；这两种都在语音识别中广泛使用 [9]。这些特征是在级联的每个阶段对相同的音频输入提取的。这两个分类器通过模型平均集成来计算一个音频窗口是关键词的概率。由于这些分类器是在同一数据的不同表示方法上训练的，因此它们在训练过的分布区域达成一致，在其他情况下则表现得随机。这导致了误报率降低。在我们的

系统中，两者都是从 30 毫秒的帧中提取的，步长为 10 毫秒，并且每个 500 毫秒的例子都进行了拼接。

3.3 级联

级联分类器常用于处理高度不对称的分类问题；它们通过在许多实际机器学习解决方案中的应用而广受欢迎；Viola-Jones 人脸检测器是一个著名的例子 [10]。

级联是一种基于多个复杂度逐渐增加的分类器串联的集成学习实例。每个分类器都在未被前一个分类器过滤掉的样本上进行训练。为每个分类器的输出设定阈值，使得其中一部分负样本（非关键词）能够正确分类。剩余的样本要么是正样本（关键词），要么是难以处理的负样本。为了训练级联中的下一个分类器，我们在一个保证不包含目标关键词实例的大音频库上运行之前的所有级联分类器，并将发现的误报作为难例负样本使用。这使得后续级联分类器能够专注于前一个分类器混淆的负样本。随着后续分类器的应用，正负数据的比例更加对称。该技术的一个缺点是，随着级联性能提高，提取用于训练后续分类器的难例需要更长时间。在我们的实验中，第一个级联分类器使用每 1 个正样本对应 100 个负样本的比例进行训练。第二个和第三个级联分类器维持每个正样本对应 2 个负样本的比例。推理时，级联分类器提供了一种方法，在非关键词窗口上提前终止后续关键词检测计算。

3.4 多实例学习

管道的最终阶段，如图 1(iii) 所示，聚合当前和过去级联分类器的输出以对目标关键词的触发做出最终决定。我们建模的 KWS 问题本质上是一个多实例学习（MIL）[11] 问题：关键词存在于音频信号中的某个位置，但其位置定义不精确且可能在持续时间上变化极大，因为人类发音同一词语的方式范围广泛。我们发现，“Okay Eva”的发音时长可以从 300 毫秒到 900 毫秒不等。

在 MIL 中，训练样本不是单个实例；它们以“包”形式出现，使得包中的所有示例共享一个标签 [11, 12] 。包含正窗口的包意味着该包中至少有一个窗口是正类，而负包则表示包中的所有窗口都是负类。在 MIL 中，学习必须同时确定哪些正包中的实例为正类，以及分类器的参数。在我们的案例中，一个包是由 10 毫秒步长组成的 500 毫秒窗口组。

因此，我们设计了学习过程，在特定的正向或负向窗口包中同时对所有窗口进行学习。包中所有窗口对应的输出通过“噪基或”聚合。

4 实验与结果

我们评估了短语“Okay Eva”的检测，使用每小时假阳性率与假阴性率作图。

多特征表示 仅使用一种特征表示（单独的 MFCCs 或 PLPs）的三阶段级联与同时使用两种表示的三阶段级联的 ROC 特性如图 2 所示。多表示方案改善了假阳性和假拒绝率之间的权衡。例如，在假拒绝率为大约 7%（0.07）时，MFCCs 的每小时假阳性率为 1.2；PLPs 的每小时假阳性率为 1.0；多表示方案的每小时假阳性率为 0.55。

级联 我们训练了一级、两级和三级级联分类器以进行比较。由于第一个分类器主要是在负例上进行训练的，因此需要计算每个级联中第一个分类器输出概率的阈值，以保证大多数正例不会被过滤掉。各个级联分类器阶段的 ROC 特征如图 3 所示。随着更多分类器的级联，ROC 特征显著提升。

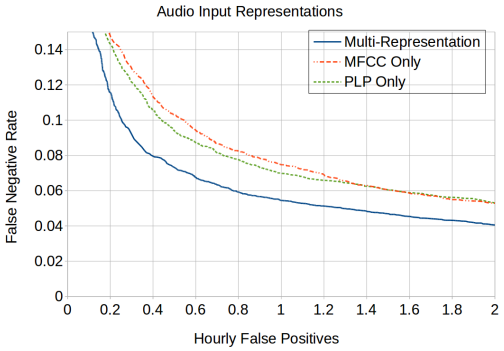

图 2: 一幅展示使用 PLPs、MFCCs 和多表示模型效果的图表。

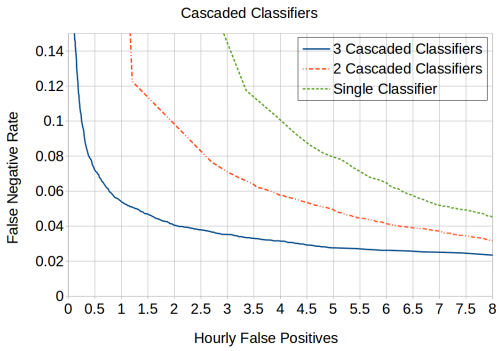

图 3: 显示使用级联分类器效果的图表

5 结论

我们在窄带音频中显著提升了关键词 spotting 的性能，同时最小化了计算资源的使用。通过结合多种特征表示和三个级联分类器，我们将在 5%假阴性率下的每小时误报次数从 8 次减少到 1.2 次，减少了 85%。尽管由于数据集的不同它们无法直接比较，但在窄带 8kHz 音频上我们的系统表现优于谷歌的 DNN 系统 [5] 在宽带 16kHz 音频上的表现。

参考文献

[1] O. Tawakol, “Introducing eva by voicera,” Feb 2017, (Accessed 29-Oct-2017). [Online]. Available: https://www.voicera.com/introducing-eva-by-workfit/

[2] L. Gallardo, *Human and Automatic Speaker Recognition over Telecommunication Channels*, ser. T-Labs Series in Telecommunication Services. Springer Singapore, 2015.

[3] S. Voran, “Listener ratings of speech passbands,” in *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, Sep 1997, pp. 81–82.

[4] S. Möller, F. Köster, L. F. Gallardo, and M. Wagner, “Comparison of transmission quality dimensions of narrowband, wideband, and super-wideband speech channels,” in *2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec 2014, pp. 1–6.

[5] T. N. Sainath and C. Parada, “Convolutional neural networks for small-footprint keyword spotting,” in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 1478–1482.

[6] G. Chen, C. Parada, and G. Heigold, “Small-footprint keyword spotting using deep neural networks,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.

[7] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, “Continuous hidden markov modeling for speaker-independent word spotting,” in *International Conference on Acoustics, Speech, and Signal Processing,*, May 1989, pp. 627–630 vol.1.

[8] S. Team, “Hey siri: An on-device dnn-powered voice trigger for apple’s personal assistant - apple,” Oct 2017, (Accessed 29-Oct-2017). [Online]. Available: https://machinelearning.apple.com/2017/10/01/hey-siri.html

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, 2002.

[10] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[11] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.