关键词识别研究计划

林正青[1)] 赵泳豪[2)] 罗天辰[3)]

[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)
[3)](南开大学 计算机学院, 天津市 300350)

摘 要 本文主要介绍了我们组基于 KWS 任务定义的简化任务和接下来的工作安排。我们选取有可参考代码的论文，主要涉及注意力机制、卷积神经网络及其变体和残差网络，我们仅在模型的软件层面进行复现和改进，同时保留对轻量级模型和噪音鲁棒性的要求，评价指标是在谷歌语音命令数据集上的关键词预测准确率。最后我们提出了打算尝试的创新工作和技术路线。

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

A Research Proposal about Keyword Spotting

LIN Zhengqing[1)] ZHAO Yonghao[2)] LUO Tianchen[3)]

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

Abstract This paper mainly introduces the simplified tasks defined by our group based on the KWS task and the following work arrangements. We select papers with reference code, mainly involving attention mechanism, convolutional neural network and its variants and residual network, we only reproduce and improve the software level of the model, while retaining the light-weight model and noise Robustness is required, and the evaluation metric is the keyword prediction accuracy on the Google voice command dataset. Finally, we propose innovative work and technical routes that we intend to try.

Key words Keyword Spotting; Convolutional Neural Network; Google Speech Command Dataset; Recurrent Neural Network; Residual Network; Attention Mechanism

1 问题定义

我们组计划探索语音命令词识别任务的解决方案。我们定义该任务为关键词识别（Keyword Spotting，KWS）任务的简化版本：目标是使用轻量级模型在谷歌语音命令数据集上取得较高的预测准确率，同时尽可能减少背景噪声或不相关语音导致的误报。相比于 KWS 任务，我们保留了对模型大小和噪音鲁棒性的要求，而仅对模型所在的软件层面进行设计和改进，不使用设备信息，也不考虑模型与硬件的结合。

收稿日期：2022-11-23；修改日期：2022-11-23 林正青，女，2000 年生，本科学历，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: xxx. 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: xxx. 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: xxx.
第 1 作者手机号码：xxx，E-mail: xxx

需要指出的是，我们考虑到数据增强是一个重要的预处理步骤，出于学习目的便保留了抵抗噪音干扰这一要求，参考论文中使用的补充数据集和方法。

2 国内外研究现状

我们定义问题的解决方案主要参考了 KWS 任务。KWS 任务中，最初主要使用 HMM 模型，在这些方案里 DNN 只作为一个获取先验概率的环节。相比传统 HMM 方法，近年出现的完全基于神经网络的端到端模型在预测性能上获得了很大提升，而代价是较大的计算和存储开销。由于使用场景为低资源移动设备，对模型轻量性的要求较高，历来研究多使用 CNN、变种 CNN 或 CRNN（CNN 与 RNN 的结合），随着新型网络的提出，又出现加入 ResNet 块或 Attention 机制的方法。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，准确率达 97% 以上，如[1]。KWS 系统运行时不断监听外界声音输入并寻找关键词，因而大部分输入接近静音，且预测时会受到背景噪声和不相关语音这两类噪音的干扰。对于噪音，除了使用滤波技术，还可以在训练模型时应用数据增强技术提升鲁棒性。此外用户自定义关键词也是一个发展中的研究方向。

3 计划复现论文

综合考虑模型先进性、组员的学习期望以及在有限时间内复现的难度，我们选择的论文大部分有开源代码以供参考。组内标准是无论基于已有代码还是完全自己复现，必须写出清晰的注释和文档以体现自己在代码实现、设计思路上的理解。

3.1 MatchboxNet[1]

该模型在 30 余种关键词分类下，使用 93K 参数实现了约 97.48% 的准确率，是现阶段容量最小、准确率最高的模型之一。模型由若干卷积模块及子模块构成，使用的 1D 时间-通道可分卷积（1D Time-Channel Separable Convolutions）网络结构基于 2019 年提出的 QuartzNet 网络，相比分组卷积参数量还要大大减少。文献采用的其他预处理策略包括 MFCC 滤波、分帧、使用 SpecAugment 和 SpecCutout 方法进行数据增强等（两者均有论文对应）。参考代码来自一些 github 用户的复现版本。

3.2 Streaming keyword spotting on mobile devices

对于本文主要复现 MHAtt-RNN（Multi-head Attention+GRU）模型。此外本文设计了一种自动化流式推理（streaming inference）模块，不影响模型训练，在推理环节进入流式推理模式，每隔几十毫秒输出一次预测结果，在现实 KWS 场景中由于不知道何时捕获关键词，工程上可以采用这种方法。如果时间允许，也将复现该模块。

该文献同时有 Tensorflow 版（官方）和 Pytorch 版开源代码。

3.3 WRN

WRN 在 ResNet 的基础上，更改了基础的 Basic Block 和 Bottleneck Block 结构，减少了网络深度，增加了卷积层宽度 (卷积核数量)，相比于 ResNet 能以更少的参数量和网络深度完成相同质量的任务 (28 层网络的结果与 ResNet101 接近)。

实现从 RestNet Basic Block 到 WRN 的改变是比较简单的：首先是通过超参数 k 增加卷积核的个数；对于 Bottleneck Block 结构，除了增加卷积核宽度之外，还要在中间增加一层 Dropout。

本方法的一个参考代码来自 2018 年 Kaggle 比赛获胜者，此方法现在排名第 10，也是效果最好的之一，不过是完全基于最初提出的 WRN[2]，没有直接针对此任务撰写的论文。因此本文就将原始论文作为本方法复现的篇目。

此外，我们拟参考前人工作对 WRN 网络作进一步改进，包括但不限于：增大卷积核尺寸从而提升网络获得更加全局的特征信息的能力；加深网络宽度从而提升网络收集全局特征信息的能力；尝试在 WRN 结构的基础上结合 self-attention 从而更好的收集全局特征信息。

4 创新想法和技术路线

除了对于原模型的调参、测试和组合优化，我们组还提出了将另外两种新模型应用于本次任务的想法，简述如下：

4.1 Conformer

Conformer 是 Google 在 2020 年提出的语音识别模型，基于 Transformer 改进而来。Transformer 结构适合于长序列，self-attention 机制提取全局信息，卷积则更适合于提取局部特征，Conformer 模型将这两种结构进行了结合，将卷积作用于 self-attention 层之后，使得模型在 self-attention 关注全局信息的同时也可通过卷积提取局部信息。卷积和 self-attention 的结合同时提升了模型在长期序列和局部特征上的效果，增加了模型的泛化性。该模型在语音识别数据集 LibriSpeech 数据集上取得了 SOTA 的结果，证明了其结合的有效性。

4.2 ContextNet

该模型主要基于卷积神经网络实现。和 Conformer 类似，该模型也考虑到了卷积由于感受野较小的局限性——卷积不能很好地感受全局信息，对于卷积神经网络需要很大深度才能将两个相距较远的数据的信息进行聚合，特别是对于语音识别/翻译等领域的长序列输入而言。

该论文证明了即使没有 self-attention 结构，只要卷积神经网络利用好全局信息，依然能发挥出不错的效果，在此基础上或许可以沿着全局信息的提取方式这条路线继续前进，寻求更优方案。

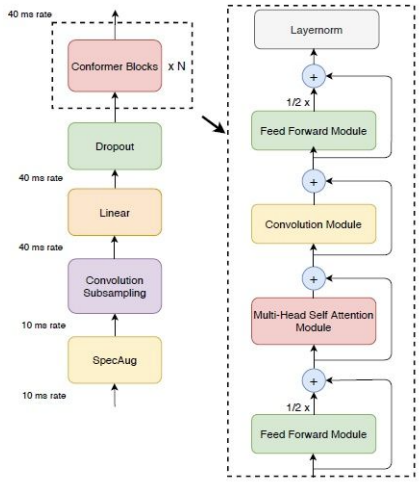

图 1 论文中给出了模型的结构，观察 Conformer 块结构，可以发现在每个 self-attention 结构之后都增加了卷积层进行局部信息提取。

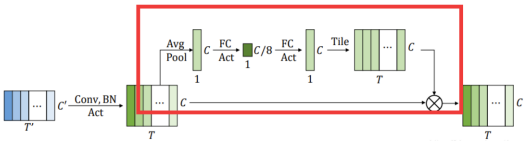

图 2 模型中使用了 SE 结构，利用卷积将一个序列的特征向量转换为一个作用于全局的上下文信息向量，然后将上下文信息向量作用于每个原始的特征向量，通过该过程使得每个特征向量都具备一定的全局信息，从而一定程度上解决了卷积神经网络对于全局信息感受弱的缺点。

5 组内成员分工

林正青：复现 MatchboxNet 和 MHAtt-RNN，搜集论文，撰写本文摘要、第一、二章和第三章 3.1、3.2 节。

赵泳豪：复现 WRN 并进行改进，进行 Conformer 模型和 ContextNet 模型的测试，撰写本文第三章 3.3 节和第四章。

罗天辰：暂未承担任何工作。

参考文献

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

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

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

[4] KRIMAN S, BELIAEV S, GINSBURG B, et al. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2020: 6124-6128.

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

[6] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.

[7] HAN W, ZHANG Z, ZHANG Y, et al. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context[J]. arXiv preprint arXiv:2005.03191, 2020.

LIN Zhengqing, Bachelor. 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.