报告题目：毕业设计调研（一）

任永文

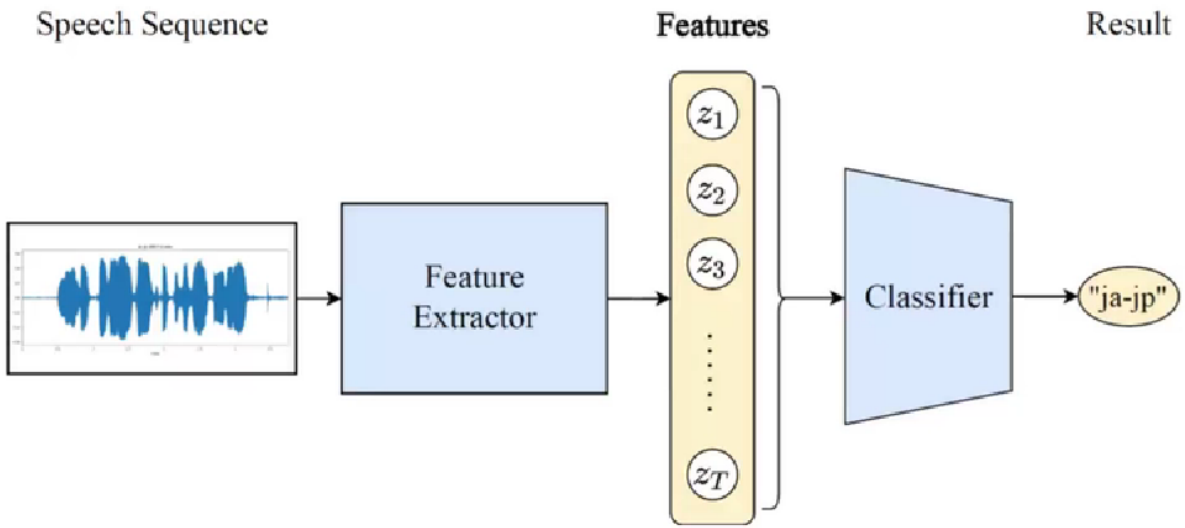
2022 年 11 月 20 日

目录

概述

工作流程

提取语音片段的声学特征，经过注意力池化层和全连接层进行语种的分类判决



概述

非深度学习

高斯混合模型-通用背景模型（GMM-UBM）

高斯混合模型-支持向量机（GMM-SVM）

基于 GMM 提取 i-vector 特征

将每条音频的 GMM 超向量映射为含有音频显著特征的低维向量，这个低维向量即为 i-vector

深度学习

用增加了瓶颈层的神经网络（BN-DNN）提取 ivector 特征

对声学特征进行多层非线性映射和降维压缩，得到鲁棒性更强的高层抽象特征

用延时神经网络（TDNN）提取 x-vector 特征

通过将不定长的语音片段映射到固定维度的 embedding，这个 embedding 即为 x-vector 特征

用（Extended-TDNN）提取 x-vector 特征

Extended-TDNN 网络拓展了时间上下文，并加入了 Dense 层，增加了网络深度

用（ECAPA-TDNN 网络）提取 x-vector 特征

采用自注意力机制和多层聚合等增强方法，进一步拓展了时间上下文，并关注到全局属性

TDNN 网络

标准的 TDNN 网络由帧级别层、统计池化层和段级别层组成

1. 帧级别层为 5 层的时延网络结构，处理语音的帧级别特征
2. 统计池化层对每一条语句的帧级别特征计算均值 和标准差，得到整条语句的全局特征
3. 段级别层考虑了部分语句的时序结构信息，由两层全连接层组成，分别提取不同的 x-vector 特征

自监督语音预训练模型

- 用前文预测当前及后文信息
- 用前后文预测当前信息
- 随机 mask 一些帧并预测

用前文预测当前及后文信息

对比预测编码（CPC）

依赖噪声对比估计训练模型，训练编码器提取输入到上下文网络的特征，在输出端进行正例和负例的鉴别性学习，优化网络参数

自回归预测编码（APC）

生成性模型，在训练过程中预测重建未来语音片段的频谱

VQ-APC：引入矢量量化层，把连续的特征映射到离散的 token 上

模型基于 LSTM 或者 GRU 模块来实现

用前后文预测当前信息

非自回归预测编码（NPC）

只需要输入被 mask 掉的帧前后的一些帧进行隐蔽重建，而不是进行预测

卷积块可以用感受野限制信息的前向传递过程，保证重建使用的信息来自周围的前后文

随机 mask 一些帧并预测

wav2vec 系列

VQ-wav2vec

原始音频片段输入到 CNN 层-VQ 层-CNN 进行特征的抽象。用 CNN 层的输出 Ci 预测未来某时刻 VQ 的输出 Zi+k，计算对比损失，训练 VQ 模型。将 VQ 层的输出作为后面 BERT 的输入，随机 mask 掉一些帧并预测，训练一个 BERT 模型。

wav2vec2.0

音频分段输入 CNN 层提取特征，后一方面输入到 VQ 层，另一方面随机 mask 掉一些帧输入到 Transformer 层提取特征。

Mockingjay&Audio Albert&TERA

基于 transformer 模型

数据集

AP17-OLR

东方语种识别竞赛提供的 10 种不同语言数据集 AP17-OLR，10 种语言分别为日语、韩语和哈萨克语（时长分别为 5.8 h、5.9 h 和 5.4 h)；粤语、普通话、印度尼西亚语（时长分别为 7.7 h、7.6 h 和 7.5 h)；越南语和俄语（时长分别为 8.4 h 和 9.9 h)，藏语和维吾尔语（时长均为 10 h)。每个语种的语音采样频率为 16 kHz。

概述

设计构想

ECAPA-TDNN + NPC/wav2vec

主任务采用 ECAPA-TDNN 网络模型，辅助任务采用改进的 wav2vec 网络模型，以帧级特征作为输入进行对比预测学习。

时间安排

- 12 月：打基础——李宏毅深度学习与人类语言处理
- 1 月：提高——阅读论文构思及创新点
- 2 月：实现——掌握 pytorch 的使用并实现代码
- 3 月：收尾——完成论文

For Further Reading I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.