浙江大学软件学院夏令营

智能语音计算

真伪语音鉴别算法评测

实验报告

姓名：孙夏恩

基于特征级联与 BinaryFocalLoss 卷积神经网络

语音真伪鉴别算法实验

1. 实验背景

本次夏令营任务语音片段真伪鉴别，根据任务给予的包含真伪两种类别的语音片段训练集，提取语音特征并训练模型，最终在评测数据集上进行判别结果测试，实验预期为设计鲁棒性良好的鉴别模型。

2. 实验环境

实验环境：Anaconda (python 3.7)，jupyterLab（2 核 4G 服务器上运行）

核心库依赖：tensorflow，keras，numpy，librosa，spafe

由于本次实验为自行搭建的算法 pipeline，为研读了相关论文后对文章提及方法做了一定的融合的成果，最终编码呈现，故整个过程主要以探索式形式进行，因此最终整理后，呈现结果为一份附带各模块功能作用说明和代码注释的 ipynb 文件，在本次实验报告中代码相关截图较少，可以在 notebook 中查看。

3. 实验过程

1. 算法 PipeLine 结构示意图

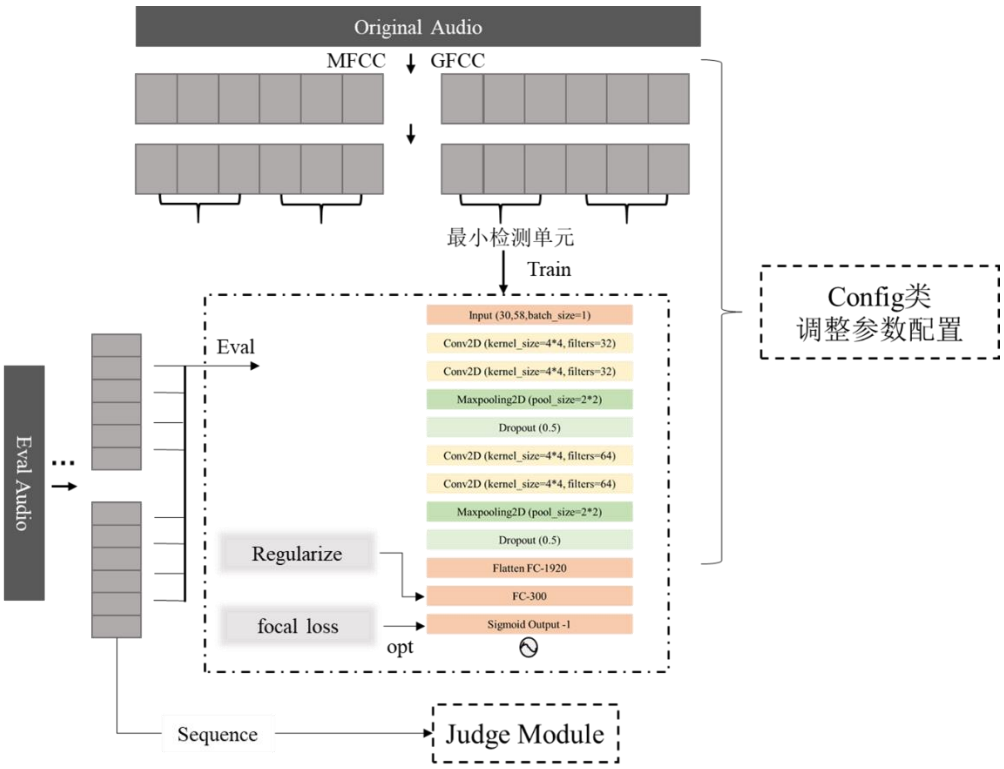

图 1 算法整体 Pipeline

2. 设置检测单元（**librosa,numpy 库**）

如图 2 语音时长分布密度图中显示，针对输入的语音片段样本，各样本语音长度差异明显，在特征提取得到的特征维度受此影响也会有较大差异，受制于对应的帧组合长度不统一。平均的时间长度在 3.106 秒。

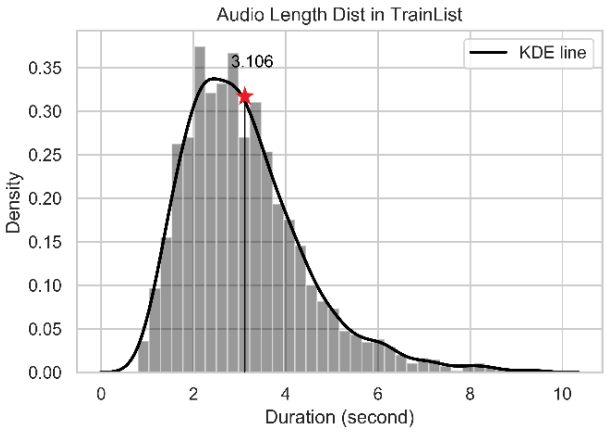

图 2 语音时长分布密度图

对于机器学习相关方法而言，输入维度需要统一，本文使用检测单元概念，将输入样本长度标准化。对以下思路进行实验：

1. 在原始语音时长视角 Duration 进行滑窗分段，即对原有语音片段进行切割，然后进入特征提取模块。
2. 在原始语音时长视角 Duration 进行滑窗分段，即对原有语音片段进行带重叠分割，然后进入特征提取模块。
3. 在 MFCC 特征提取后的帧序列，映射寻找 Duration 对应的帧长，进行分段。

表格 1 最小检测单元提取实验

方法	val 平均准确率
A．原始语音/0.6s	92.82
B．原始语音/0.5s	91.32
C．原始语音/0.6s+overlap0.1s	90.10
D．MFCC 提取帧序列后切割 +0.6s 映射	**93.16**

实验结果显示，在 MFCC 特征提取后的帧序列中进行分割可以更好的利用和表达分段特征。初步推测这很大程度上是得益于 MFCC 在特征提取的有重叠采样和加汉明窗机制，一定程度上缓解了两段相邻语音在交界处的边缘信息丢失情况。至于手动重叠采样的效果不佳，推测是由于和 MFCC 的分帧机制结合后存在一定程度上的过采样情况出

现，导致实际训练样本出现不平衡情况，因此最终本次实验选用 D 方法：**MFCC 提取帧序列后切割/0.6s 映射。**流程示意图 3 最小检测单元提取流程：

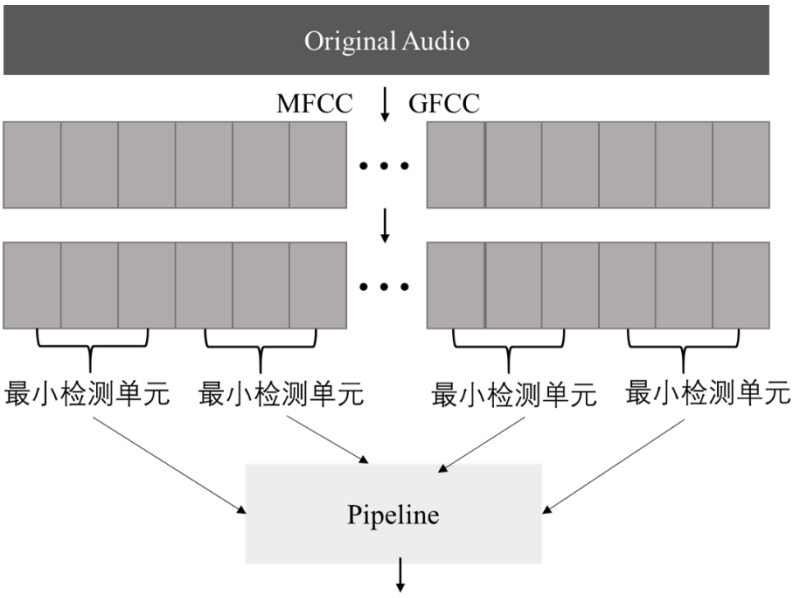

图 3 最小检测单元提取流程

补充前置预处理机制（librosa 库）：针对 OriginalAudio 层的数据，先利用 librosa 进行一次静寂声音裁剪，做一次初步过滤，提取有效信息。

补充 Padding 机制：针对特殊情况如原始语音长度小于 0.6s 的片段，采取 padding 策略，在语音特征向量中的每一特征维补齐值为 0 的特征参数，以维持一致的帧序列长度。

3. SMOTE 样本增广尝试（**imblance 库**）

表格 2 数据正负例分布

数据集\类别	Spoof 标签	Bonafide 标签	正例占比
Train（2000 total）	1779	**221**	11.05%
Dev（2500 total）	2242	**258**	10.32%

根据表格 2 数据正负例分布中数据分布显示，数据集正负样本存在明显的不平衡情况。从数据层面可以通过机的增大可获得的少数类的样本数量（过采样思想）来实现少类别数据的增广和平衡化。本次设计使用的 SMOTE 算法可以有效考虑原始样本分布，结合 K 近邻算法，在高维空间中，样本所代表的点与其相近特征点之间进行插值，达到数据合成目的。具体步骤如下：

1. **扩容参数**：确定扩容数量 S 并计算每个样本点需要相应引导的扩容样本量 N，通过配置，本次通过修改参数 sampling_strategy={1: 4000}调整比例

2. **度量（KNN）**：确定每个样本点的 K 个最近邻样本点为特征距离度量依据
3. **生成（线性插值法）**随机的选择 K 中的 N 个邻近点进行差值乘上一个处于[0,1]范围的阈值
4. 此操作需要在特征维度统一之后进行，故实际执行顺序在特征提取之后。在最终的实验过后，发现此方法进行增广数据，会有较严重的过拟合现象，因此，最后使用 Focal Loss 作为最终实验方法，详见 7 二分类损失函数接入 Focal Loss（参考论文：Focal Loss for Dense Object Detection. ICCV2017）。

4. MFCC 与 GFCC 特征提取，特征融合（**spafe 库**）

经过 Librosa 语音库装载得到的语音为 1 维数组，长度与采样率和语音的时间长度有关。本次实验训练集语音片段原始采样率为 16000HZ，需要维持各输入采样率不变，实验中进行一次 16000HZ 的重采样，确保输入的同样时间间隔内的语音片段表示成采样点数组时，具有相同的数组长度。

Spafe 库中的 MFCC 特征提取中有关产生分帧的参数设定研究结果表明，依据生物发声原理，每帧声音长度默认 0.025s，两个相邻帧之间的滑动间隔默认是 0.01s，MFCC 库函数中设置如下：

- win_len = 0.025
- win_hop = 0.01

因此，16000HZ 采样率，对应分帧为帧长=400，帧移=160。实验计算 0.6s 语音片段的固定帧长 x 方法如下：

$$0.6 * 16000 \geq 160 * x + 400$$

$$x \geq 57.5$$

因为 x 为整数，所以实验中 x 取 58，计算得到的 MFCC 最小检测单元的 numpy 数组的形状为（58 ，n_mfcc），n_mfcc 为待调整特征通道数量。

对于一帧有 512 维采样点的数据，经过 MFCC 可以提取出最重要的 40 维的特征，包含一维的能量特征，以及 3 组各 13 维的特征，可以分别对应 MFCC 系数、一阶差分参数、二阶差分参数。对于 58 维的帧序列，每一帧对应拥有若干特征系数，这个系数的维度与 58 不能相差过远，不然影响到后续的卷积操作。根据参考论文[1]中提到的 GFCC 的特征有效性在 21 组系数之后会有明显降低的先验知识，以及利用 15 组系数能够提取到较高的说话人信息的论文实验结果。因此，进行如下实验，初步考虑以级联形式合并 MFCC、GFCC 特征，提升每帧特征提取的系数数量。实验主要关注级联对于指标的正负向影响验证以及级联配比参数的微调寻优。

表格 3　n_mfcc+n_gfcc 最优参数配凑实验

N_MFCC	N_GFCC	use_energy log(C0)	平均 val_acc
12（C1-C12）	15	False	93.01
13（C1-C13）	15	False	92.32
14（C1-C14）	15	True	**93.16**
15（C1-C15）	15	True	92.97
16（C1-C16）	21	True	92.08
18（C1-C18）	21	True	92.9
39（C1-C39） （含差分）	-	True	90.7
13（C1-C13）	-	True	87.8

特征级联：从实验结果观察，选择 30 维特征融合模式为较优的特征表示，在代码中对两组特征（MFCC、GFCC）进行转置后，利用 numpy 数组的 concatenate 对两组特征进行 axis=0 方向上的级联。

Use_energy：该参数的指定对实验结果成正向，故使用对数处理过的能量分量作为第 0 维度的特征表示。

数据标准化：对于所有最小检测单元的集合进行数据标准化，计算各维度上的均值和方差，得到结果。Numpy 数组的维度特征为（sample_num，n_mfcc+n_gfcc，detection_unit_length，1），对应本实验则为（sample_num，30，58，1）。

5. 基于卷积神经网络搭建二分类模型（**keras 框架**）

在网络搭建过程中，实验有对比地参考了 LeNet 的设计方法，将图像识别的相关技巧迁移至语音任务中，利用卷积层可以更好的用来识别语音帧的频域特征。网络主要由以下五种前推模块构成。

a) Conv2D

- Kernel：采取 4*4 卷积核
- Stride：（1，1）

- Padding：Valid

b) MaxPooling2D

- Pool_size：采取 4*4 卷积核
- Stride：（1，1）
- Padding：Valid

c) DropOut

添加 Dropout 层可以有效减轻网络的过拟合现象，该层只会在训练过程中有作用。本次实验设置阈值为 0.5。

d) 全连接 Dense 层

在最后的全连接区域添加 L2 正则项 0.0001，减轻多次迭代后，过拟合对模型的影响。添加正则项可以控制这一层层参数不至于过大。

e) Relu 函数激活

Relu 激活函数在不同的参数初始化方法下使模型更容易训练。ReLU 激活函数在正区间的梯度恒为 1，受到网络初始化的影响更小。

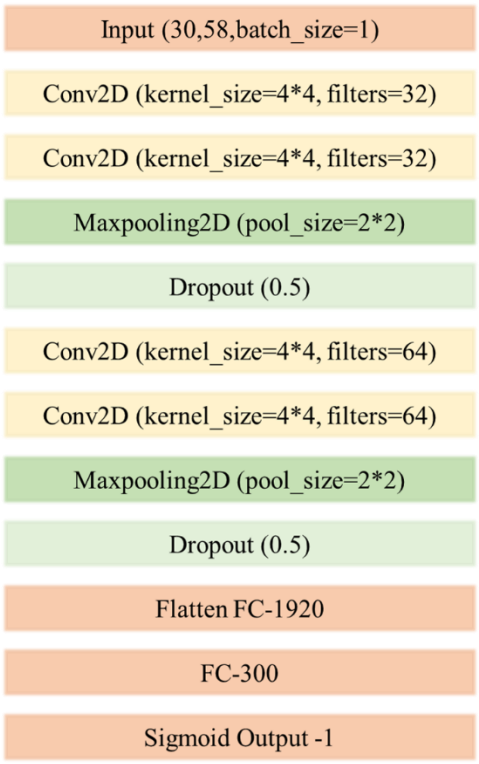

图 4 详细网络结构图

6. 训练超参数设置

a) Epoch：上限 100，设置早停，patience=10，当 10 轮迭代没有在验证集中出现 loss 降低的情况，则早停。

b) Batch_size：设置为 1，经过多次实验发现，大的 batchsize 并没有出现明显的增益，故最终仍然将 batch_size 设置为 1。

c) Optimizer：SGD，采取随机梯度下降的方法训练网络，随机梯度下降初始值学习率 lr 设置为 0.01，不设置动量。

d) Lr_Decay：采取回调函数的方式监测每次训练轮次后达到平台期间的情况，如果连续 5 轮迭代 train_loss 不继续下降，则施加学习率衰减系数 factor=0.7，减小学习率继续学习，同时设定最小学习率兜底值 min_lr=0.001。

7. 二分类损失函数接入 Focal Loss（**参考论文：Focal Loss for Dense Object Detection. ICCV2017**）

正负样本比例失衡的情况下，如果使用传统的二分类对数损失函数，最终输出层的单个神经元结果采用 Sigmoid 函数激活后得到$\sigma(x)$，y为真实样本标签（0/1），观察损失函数L_{ce}可以发现损失函数的计算对于无论是正例还是负例都是一视同仁的，损失大小根据神经元输出概率与 Sigmoid 函数的激活阈值 0.5 差值成对数相关。

$$L_{ce} = -ylog\sigma(x) - (1-y)log\sigma(-x)$$

本次实验中正例的低占比使得模型在普通二分类损失函数使用中，会更倾向于负例，正例带来的损失容易因为大规模的负例的损失的存在，受到模型的关注度降低。因此，本实验结合了图像处理目标检测领域重要的 Focal loss 损失函数L_{fl}设计，迁移场景至本次语音真伪识别的任务中。

$$L_{fl} = -y(\alpha)\sigma^{\gamma}(-x)log\sigma(x) - (1-y)(1-\alpha)\sigma^{\gamma}(x)log\sigma(-x)$$

指数γ的使用能够使得模型在倾向于学习某一类型特征时候，不再单一地去反馈多数类的损失降低。以本次背景为例，模型在针对负例学习过程中，当负例输出层判别概率越来越接近 0 时候，能够提供的损失也就越来越少，而对于神经网络学习较困难的正例，由于仍然具有较大的概率偏差，在γ的帮助下可以产生增益。至于**超参数α**的引入，目的是对于指数修正的降权处理，原论文提供了经验值[2]，因此本次损失函数的超参数采取组合$\alpha = 0.25$，$\gamma = 2$。

经训练，Epoch=38 时能够得到较好的泛化性能和拟合能力，过使用这一时刻的 Checkpoint 模型文件作为实验模型结果。最终，多次实验和训练后，受到到随机影响，实验得到的模型在验证集上的准确度可以在 0.931 附近波动，召回率达到 0.81。

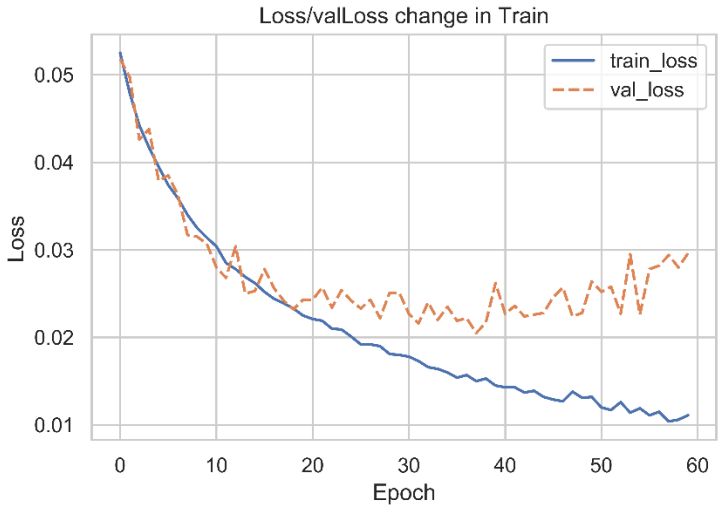

图 5 训练损失函数变化图

8. Eval 阶段原始语音数据分段检测类别判定算法

训练、测试数据经过定长分片后，每一个检测片段为原始语音的一个组成部分，同时获得一个模型输出的预测概率（标签）。对于每一条评测时期进入网络的语音，对其进行相应的分片，假设共分为 X 片段，标准化处理后，送入模型 Pipeline，得到的输出结果也为 X 组判别概率值，具体判别算法如下：

a) 设特征片段分割后的片段数量为X

b) 超过$X/2$（下取整）的片段判别概率大于 0.5，则整条语音判定为 bonafide

c) 片段判别概率大于 0.5 和小于 0.5 的数量一致，则统计X段的判别概率总和，与$0.5 * X$作比较，若大于，则整条语音判定为 bonafide

d) 除去以上两种情况的所有情况，则整条语音判定为 spoof

4. 测试结果

在 Dev 数据集上，通过模型的最终调整，准确率达到了 0.9696，同时拥有 0.9341 的精确度和 0.81 的 recall rate，作为自行搭建的算法 pipeline，我认为模型在对于过拟合方面仍然有很大的改进空间，因为极为不均衡的样本类别是网络学习的痛点，Focalloss 的优化和迁移仅仅只是是一个突破。算法的优化还可以更多结合到 MFCC 和 GFCC 上，从频域中寻找更多发现，更好利用差分特征等。

同时，在这一次 eval1 的评测中，我的算法获得了 92.58 的 acc 成绩，位列排位第二名，初步可以说明此模型的设计是有效的，在过拟合方面的控制也有一定的成效。

5. 参考文献

[1] 周萍, 沈昊, 郑凯鹏. 基于 MFCC 与 GFCC 混合特征参数的说话人识别[J]. 应用科学学报. 2019,37(01):24-32.

[2] Lin, Tsung-Yi, Priya Goyal, Ross B. Girshick, Kaiming He and Piotr Dollár. Focal Loss for Dense Object Detection[C]. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 2999-3007.

[3] 甄斌,吴玺宏,刘志敏,迟惠生. 语音识别和说话人识别中各倒谱分量的相对重要性[J]. 北京大学学报(自然科学版),2001,{4}(03):371-378.