信也科技杯 | 信也 FINVOLUTION

2024第九届信也科技杯
全球AI算法大赛

智辨真言 · 数启未来

2024 FinVolution Global Data Science Competition
Deepfake Speech Detection Challenge

信也科技杯 | 信也 FINVOLUTION

第九届信也科技杯全球AI算法大赛
语音深度鉴伪识别

分享人： Diting团队

2024 FinVolution Global Data Science Competition
Deepfake Speech Detection Challenge

2024第九届信也科技杯
全球AI算法大赛

信也科技杯 | 信也 FINVOLUTION

目录

2024第九届信也科技杯
全球AI算法大赛

团队介绍-谛听

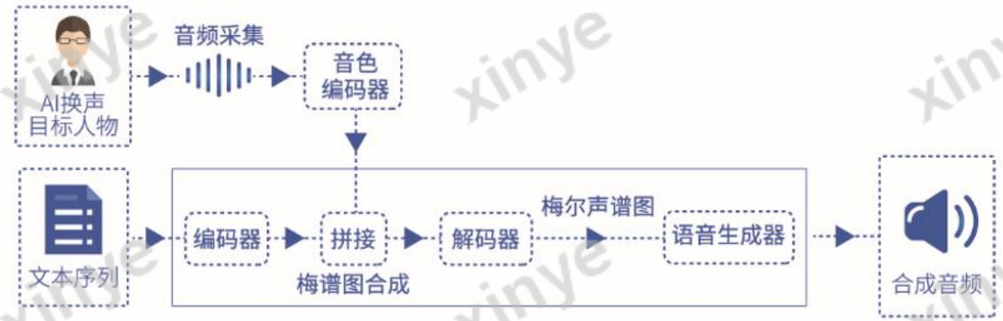
“坐地听八百，卧耳听三千”
“能辨别世间万物的声音，尤其善听人心，能顾鉴善恶，察听贤愚。”

西游记 真假美猴王 剧照

问题分析

赛题目的：

鉴别深度伪造声音，防范语音诈骗

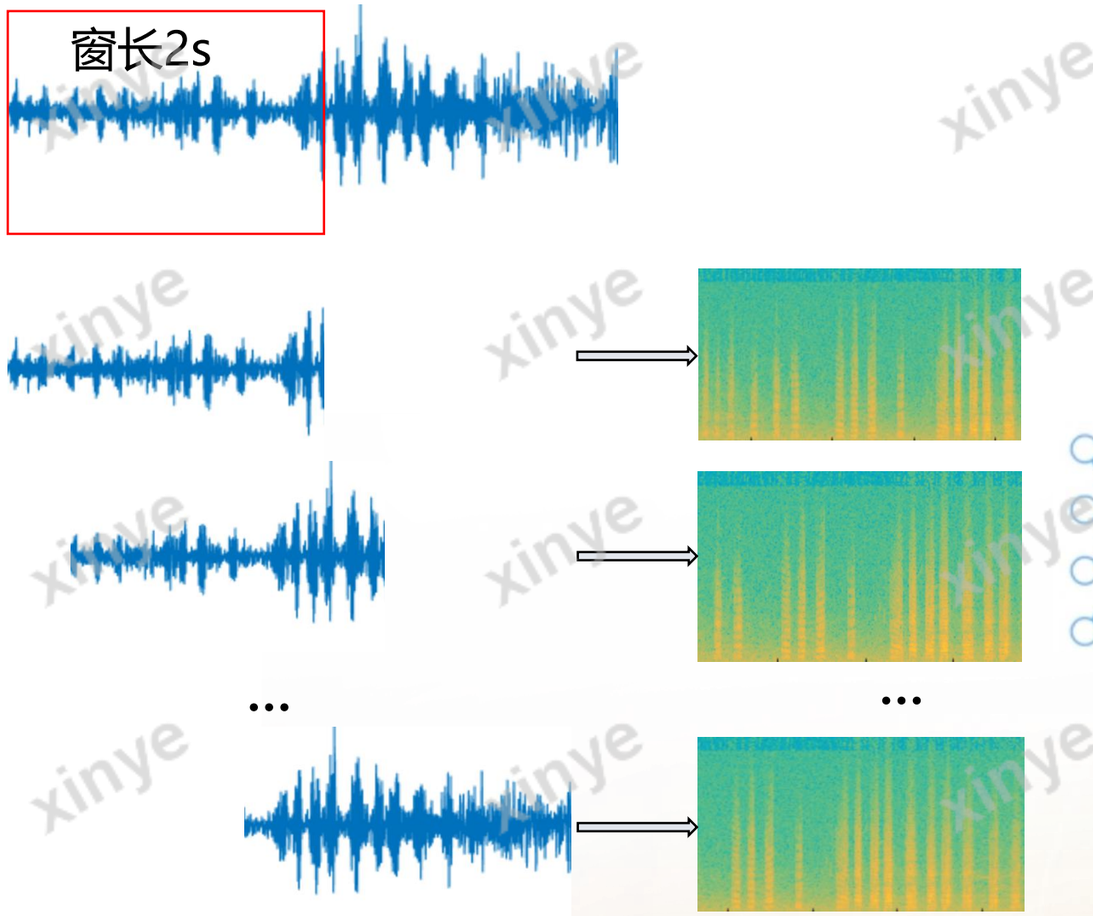


语音伪造过程

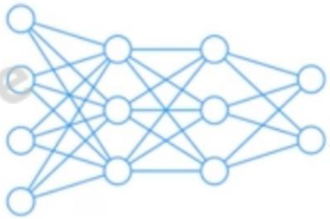
语音伪造攻击流程

Reference: https://www.jiqizhixin.com/articles/2021-04-18
https://www.bilibili.com/video/BV1rZ421s7Cs

Baseline算法分析



模型: CNN



fake score:

0.8
0.8
0.8
0.9
0.95
0.98
0.9

后处理:求均值

fake or real?

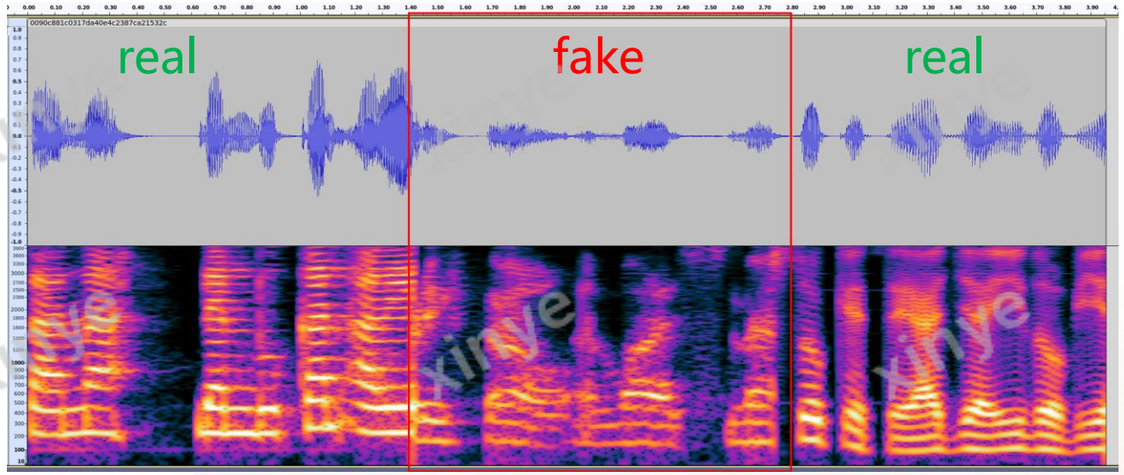
主要问题:

- 特征、模型结构即后处理较为简单

数据分析

主要挑战：

- 复赛数据与初赛数据分布差异较大：
  - **拼接：含有任一假片段即为假，后验平均方法难以检出**
  - 时长
  - 语种
  - 其它伪造方法




```
==================================================
21
outputs_softmax:
tensor([[    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.74,     0.26],
        [    1.00,     0.00],
        [    1.00,     0.00],
        [    0.99,     0.01],
        [    0.99,     0.01],
        [    0.87,     0.13],
        [    0.06,     0.94],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00],
        [    0.00,     1.00]], device='cuda:0')
outputs_softmax.shape:  torch.Size([20, 2])
softmax_mean:  tensor([[0.28, 0.72]], device='cuda:0')
mean:  tensor([[-2.53,  2.58]], device='cuda:0')
eval result:  1
==================================================
```

real

fake

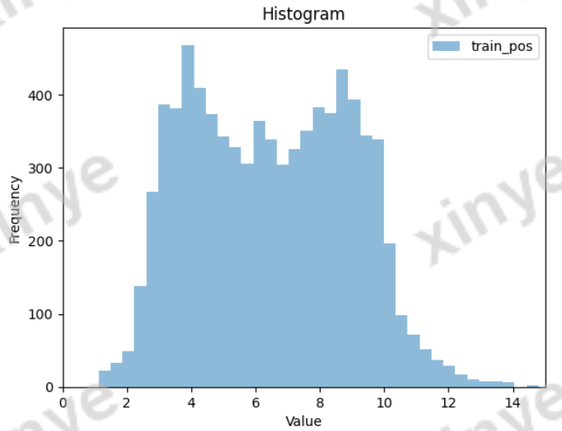
real

real

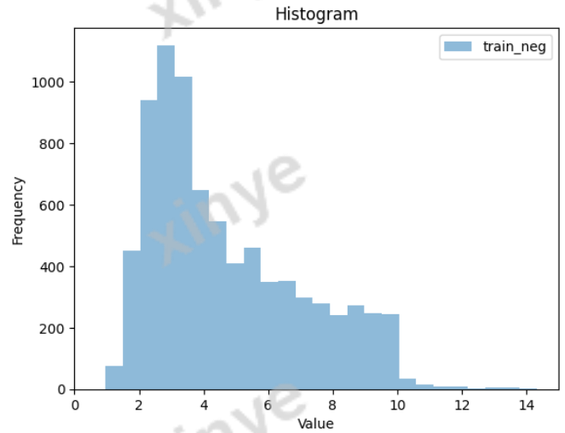
数据分析

主要挑战：

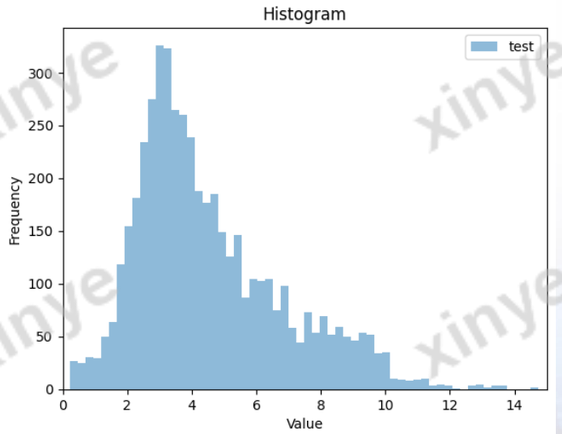
- 复赛数据与初赛数据分布差异较大：
  - 拼接
  - **时长**
  - 语种
  - 其它伪造方法



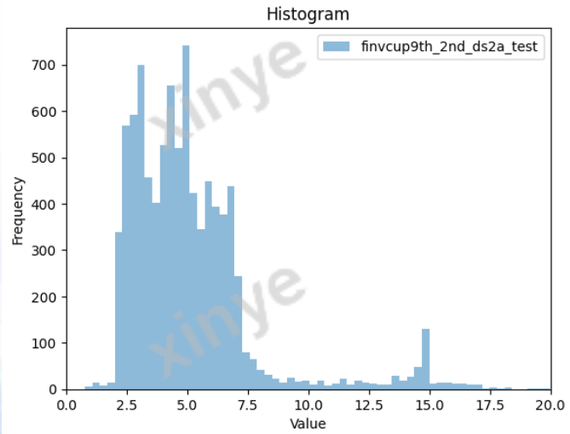
min: 1.12175
max: 23.296



min: 0.9625
max: 33.0125



min: 0.224s
max: 14.75s



min: 0.777875
max: 238.864

数据分析

主要挑战：

- 复赛数据与初赛数据分布差异较大：
  - 拼接
  - 时长
  - **语种： 包含多种语言**
  - 其它伪造方法

cnt	lang code	language	语种
16724	zh	chinese	汉语
11875	en	english	英语
1416	ja	japanese	日语
178	ko	korean	韩语
157	es	spanish	西班牙语

训练集基于 whisper large v3 语种分析结果

数据分析

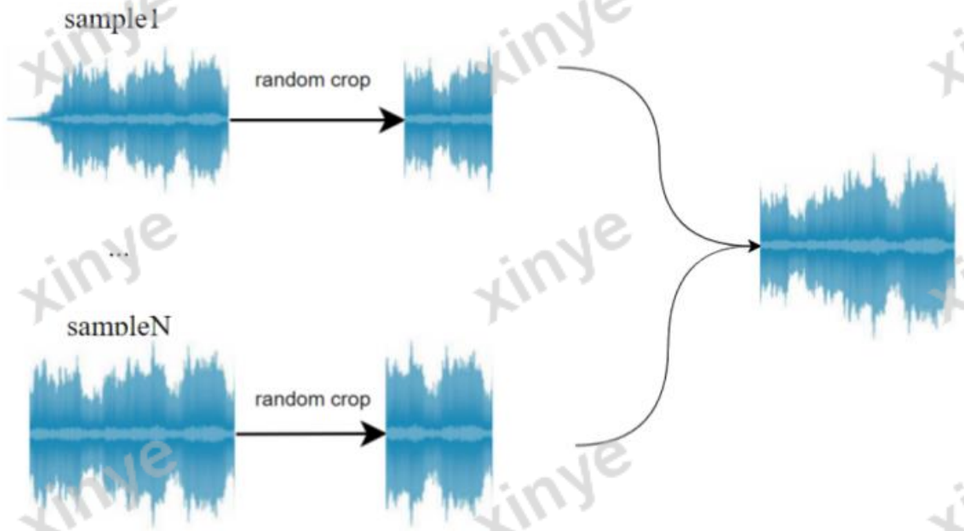
主要挑战：

- 复赛数据与初赛数据分布差异较大：
  语种
  时长
  拼接
  其它伪造方法：（考察模型泛化性）
    - 语音合成(Text-to-speech, TTS)
    - 语音转换(Voice Conversion, VC)
    - 语速扰动：变速变调、变速不变调

数据增广-动态拼接

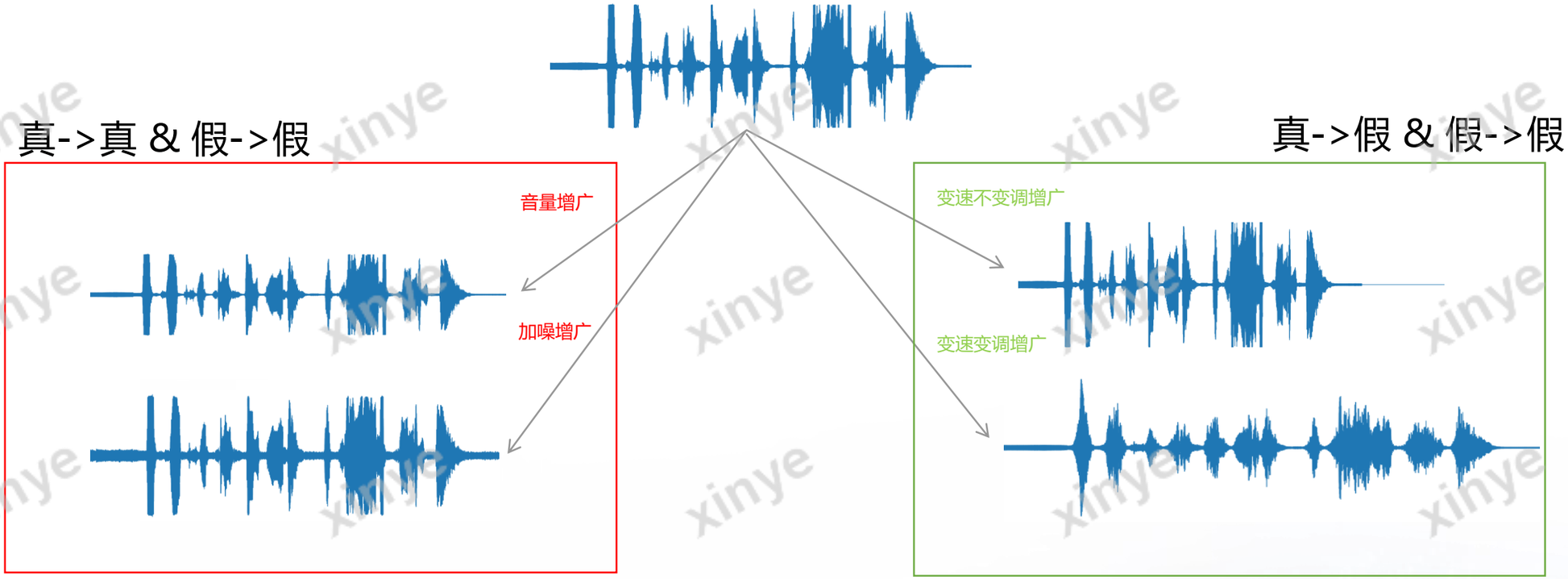
拼接结果对应类别

- 信源全真则拼接结果为真
- 任一信源为假则结果为假



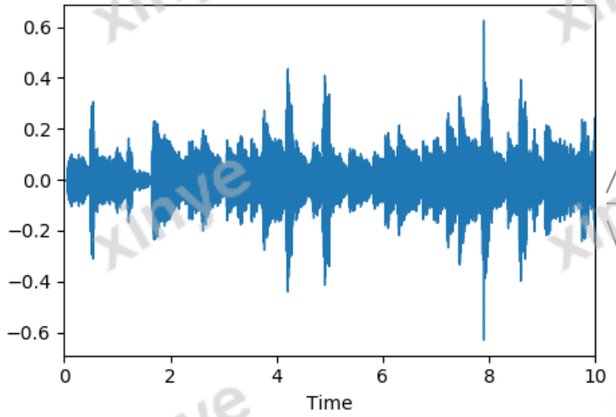
初赛	Baseline	拼接增广
F1 score	0.780662	0.798716

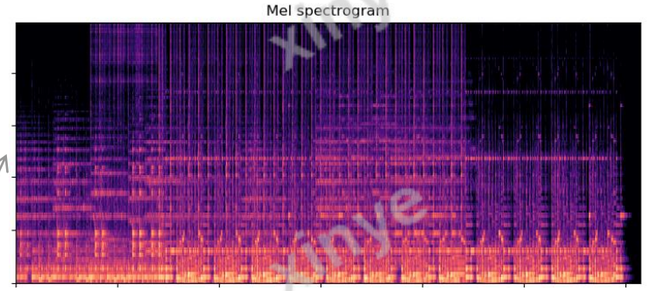
数据增广-加噪/变速/音量

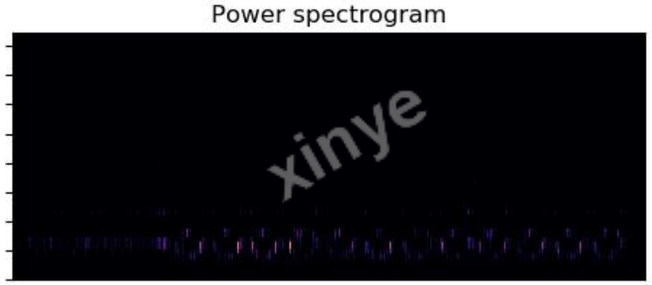


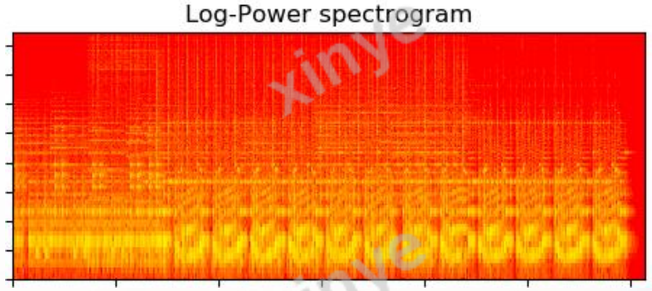
初赛	Baseline	数据增广
F1 score	0.780662	0.784903

音频特征选择









Mel-scale spectrogram:
更符合人耳听觉特性，但对高频声音相对不敏感，可能导致一些鉴伪关键细节信息的损失

Power spectrogram:
保留了所有的频率信息，对于需要捕捉细微频率变化的任务非常有用，但在噪声环境下可能不够鲁棒

Log-power spectrogram:
增加动态范围，使得声音信号的能量特性更加清晰，同时减少噪声对分析结果的影响

初赛	Ori mel spectrogram	log-power spectrogram
F1 score	0.780662	0.831576

增加开源训练数据-提高模型泛化性

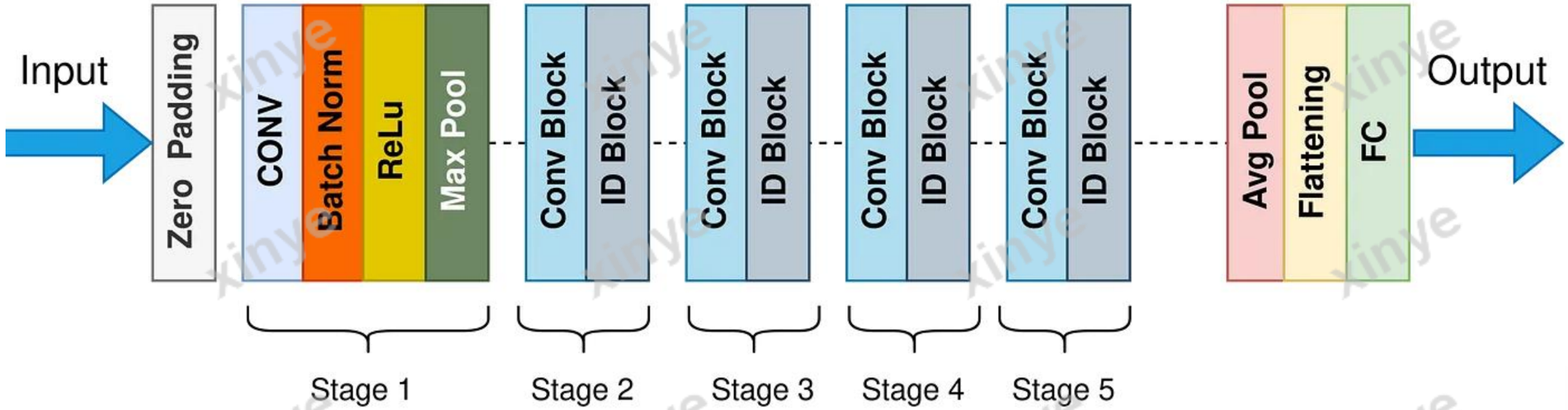
baseline训练数据量少、伪造数据种类少

Dataset	Types	Years	Language
Finvcup9th_1st_ds5_train_data	TTS	2024	UNK
CommonVoice	Real	2024	Multi-linguage(5)
In-the-Wild	TTS	2022	English
WaveFake	TTS	2021	English, Japanese
FAD	TTS	2022	Chinese
CFAD	TTS	2021	Chinese
DECRO	TTS和VC	2023	English, Chinese
MLAAD	TTS	2024	Multi-Language (23)
M-AILABS	Real	-	Multi-Language (9)
CMFD	TTS	2022	English, Chinese
FMFCC-A	TTS和VC	2021	Chinese

初赛	Baseline dataset	增加开源数据
F1 score	0.831576	0.991192

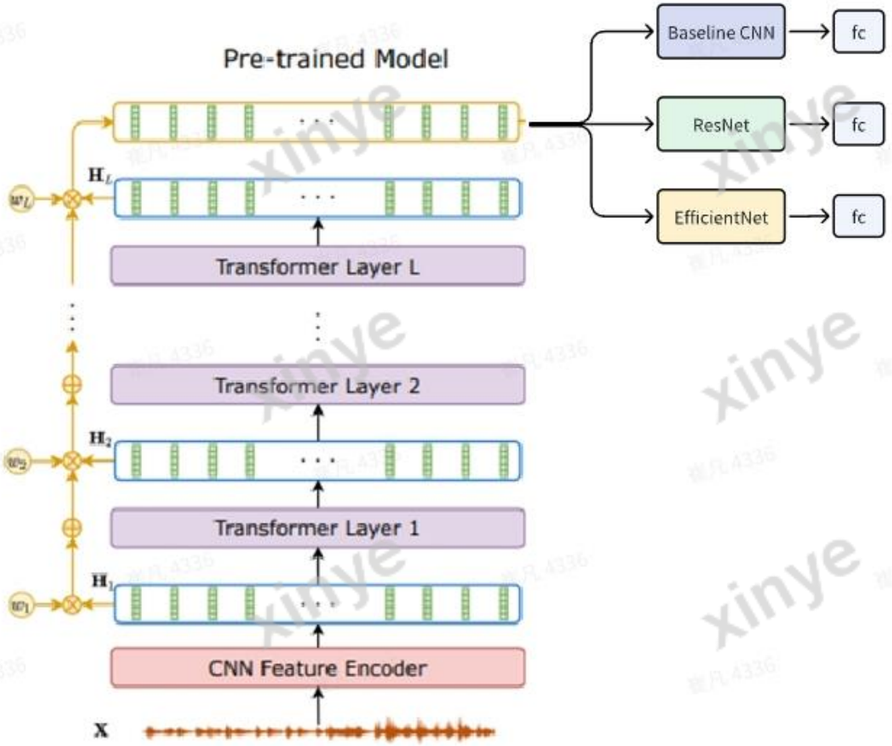
Reference: https://ai.ppdai.com/mirror/goToMirrorDetailSix?mirrorId=34

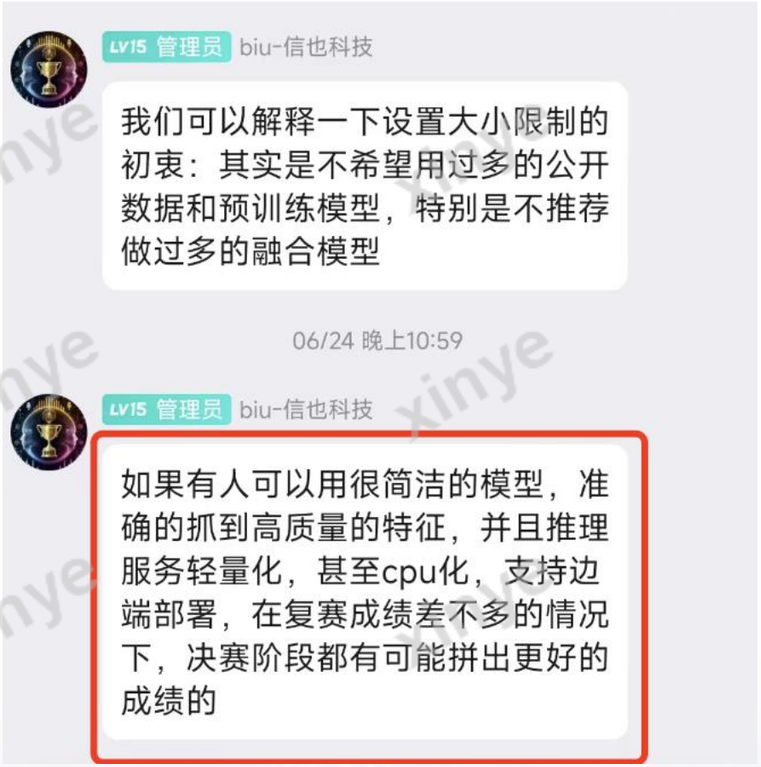
模型结构-ResNet



复赛	Baseline CNN big	ResNet
F1 score	0.668	0.723

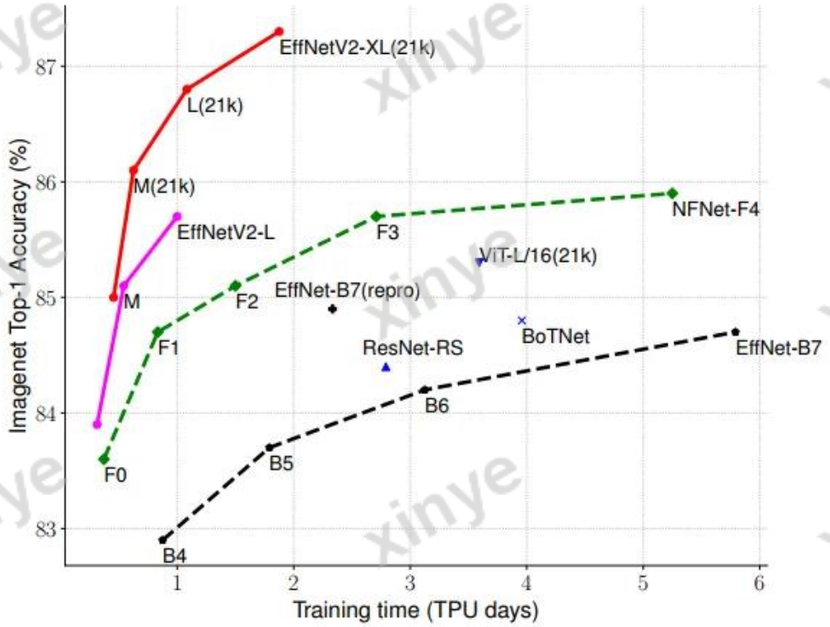
模型结构-Wav2Vec预训练

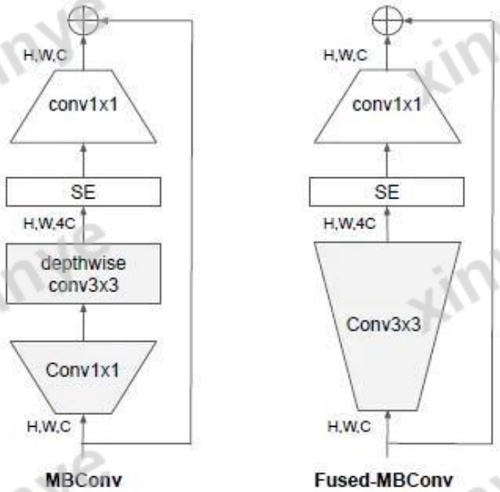




复赛	Baseline CNN big	ResNet	Wav2Vec2.0+CNN
F1 score	0.668	0.723	0.547

模型结构-EfficientNet系列





Structure of MBConv and Fused-MBConv

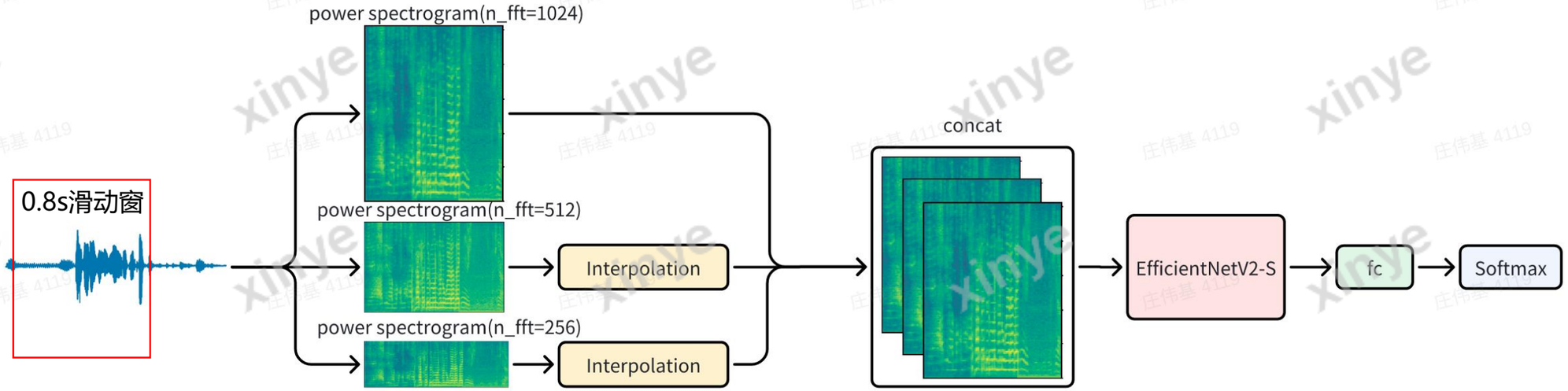
Stage	Operator	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv1x1 & Pooling & FC	-	1280	1

EfficientNetV2-S Architecture

复赛	Baseline CNN big	ResNet	EfficientNet	EfficientV2
F1 score	0.668	0.723719	0.740878	0.747714

Reference: EfficientNetV2: Smaller Models and Faster Training(https://arxiv.org/pdf/2104.00298)

模型结构-TriSpectEfficientNetV2



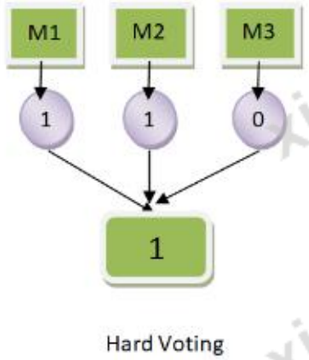
1. Test min : 0.777875s -> num_frames=80
2. Tri-Spect: 时域分辨率 VS. 频域分辨率 -> "全都要"
3. Interpolation + concat: 当成一张三维图片

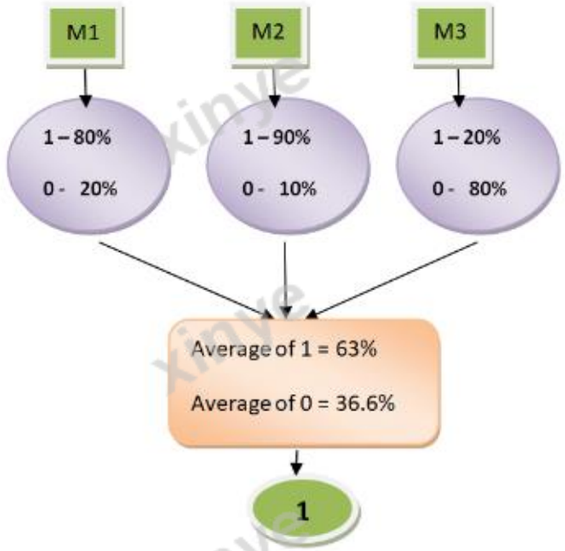
复赛	Baseline CNN big	ResNet	EfficientNet	EfficientV2	TriSpectEfficientNetV2
F1 score	0.668	0.723719	0.740878	0.747714	0.774543

后处理融合-投票

投票策略：

- 不同模型；
- 不同训练轮数；
- 投票基数；





复赛	TriSpectEfficientNetV2	TriSpectEfficientNetV2 voting
F1 score	0.774543	0.776151

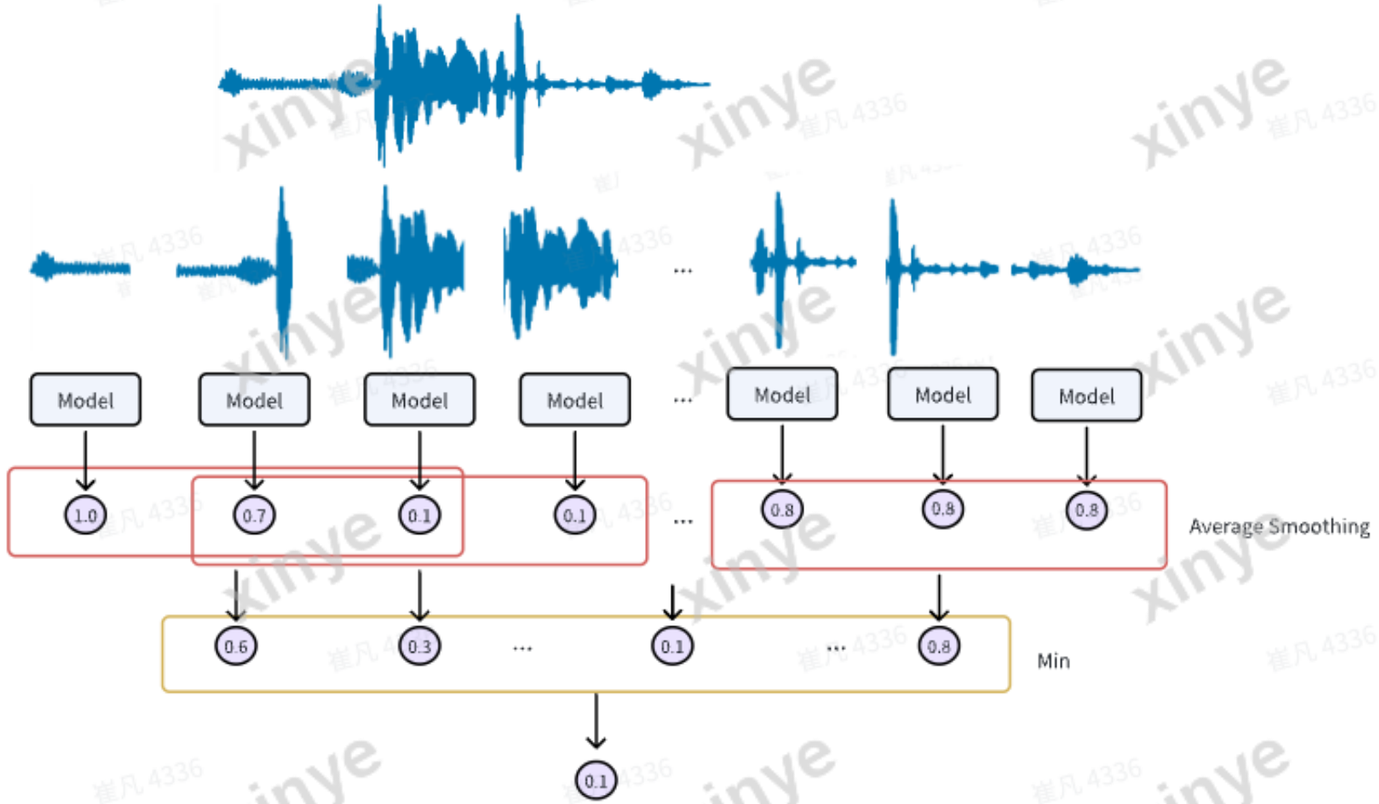
后处理-流式平滑

底层逻辑：

- 应放尽放；
- 应判尽判；

1. Smooth

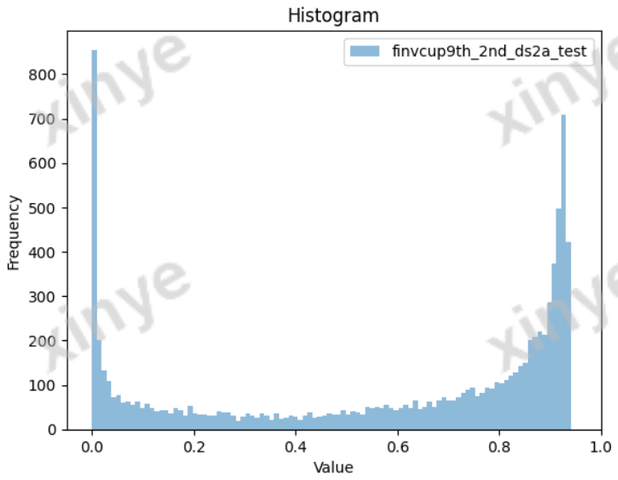
2. Mean -> Min



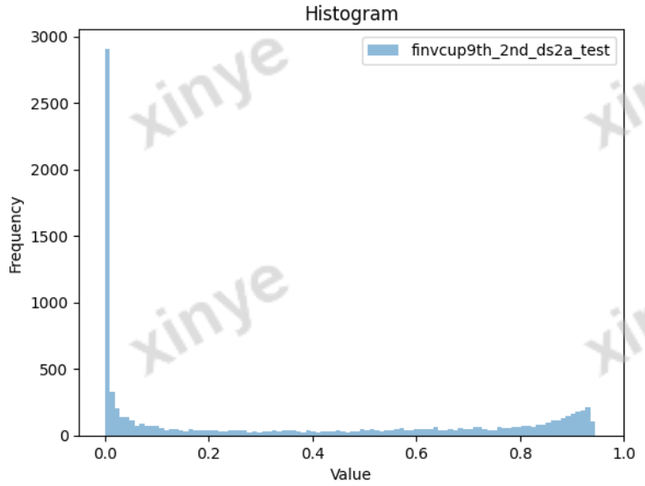
后处理-流式平滑

1. Smooth（应放尽放）：对模型的预测结果使用滑动窗口平均，减少结果中的噪声和突变，使输出更加平滑。
2. Mean -> Min （应判尽判）：防止非常确定是Fake的数据被平均为Real

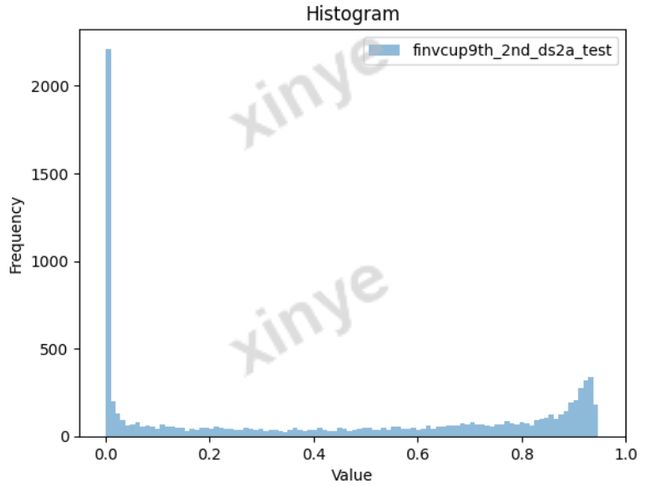
Orig mean:



Min w/o smooth



Min with smooth3

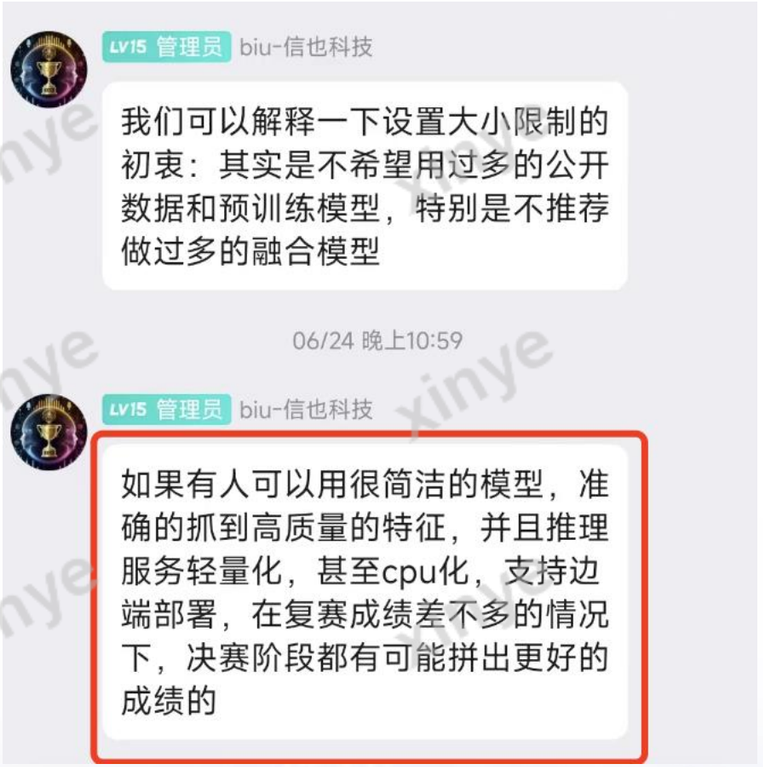


复赛	TriSpectEfficientNetV2	TriSpectEfficientNetV2-processed
F1 score	0.774543	0.799178

结果分析与总结

主要提升点：
- **数据方面：**数据增广、数据拼接、增加开源数据
- **模型方面：**TriSpect、EfficientV2
- **后处理：**模型平均、模型投票、Smooth、mean2min

无用尝试：
- **模型方面：**预训练模型
- **后处理：**模型投票



比赛感受

Simple is better !

Never give up !

信也科技杯 | 信也
FINVOLUTION FINVOLUTION

THANK YOU

2024 FinVolution Global Data Science Competition
Deepfake Speech Detection Challenge

2024第九届信也科技杯
全球AI算法大赛