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

三个臭皮匠顶一个诸葛亮

音频异常侦测：基于预训练与微调模型的ensemble探索

分享人：彭欣怡/马千里

来自：华东师范大学/Shopee

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

1. 团队介绍
2. 赛题理解
3. 整体框架
4. 关键技术
5. 应用评估
6. 总结收获

2024第九届信也科技杯
全球AI算法大赛

赛题理解

赛题描述

赛题背景

语音deepfake研究。深入工业应用场景，对伪造语音进行了重新定义。

数据描述

多语言，防止通过标签分析进行人工打标；

多来源，多种伪造方式保证算法泛化能力；

大模型，检验算法对最新伪造模型的识别效果；

混杂对抗，单个样本包含多种来源的音频。

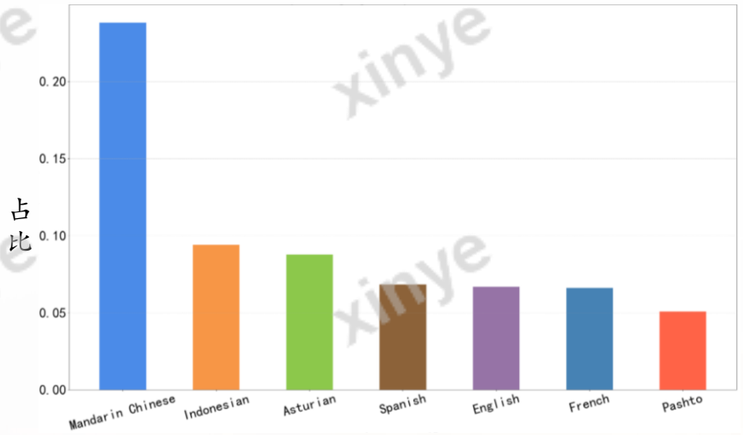
评价指标

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

* F1指标同时考虑了精确率和召回率，这两者在异常检测中都非常重要。如果只关注精确率，可能会忽略那些被错误分类为正常的异常样本；如果只关注召回率，则可能会将很多正常样本错误地分类为异常。

赛题理解

多语言



测试集语言分布

*recognized by whisper

大模型

NEW

CAMB-AI/MARS5-TTS
Text-to-Speech • Updated about 23 hours ago • ⭳7.28k • ♡ 368

fishaudio/fish-speech-1.2
Text-to-Speech • Updated 4 days ago • ⭳676 • ♡ 126

coqui/XTTS-v2
Text-to-Speech • Updated Dec 12, 2023 • ⭳870k • ♡ 1.46k

myshell-ai/OpenVoiceV2
Text-to-Speech • Updated Apr 24 • ♡ 227

microsoft/speecht5_tts
Text-to-Speech • Updated Nov 8, 2023 • ⭳234k • ♡ 571

metavoiceio/metavoice-1B-v0.1
Text-to-Speech • Updated Apr 3 • ⭳1.2k • ♡ 728

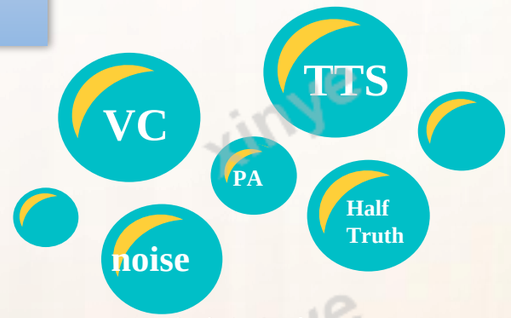
suno/bark
Text-to-Speech • Updated Oct 4, 2023 • ⭳28k • ♡ 957

facebook/mms-tts
Text-to-Speech • Updated Jul 25, 2023 • ♡ 116

近期新大模型

*from huggingface

多来源



伪造数据来源

混杂对抗

多语言混杂

多语言与噪声混杂

混杂样本案例

赛题理解

核心解法

多语言

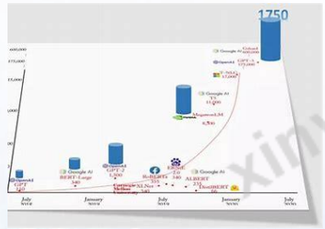
- 加入高频外语的正常语音，使模型能适应多语言；
- 选择在多语言数据预训练的模型，保证基础知识充足。

你好！Hello！

大模型

NEW

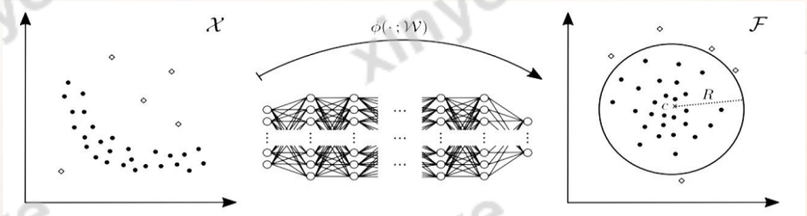
- 加入最新伪造样本(hugging face -- multi-tts)；
- 使用同源大模型进行特征提取。



模型参数趋势图

多来源

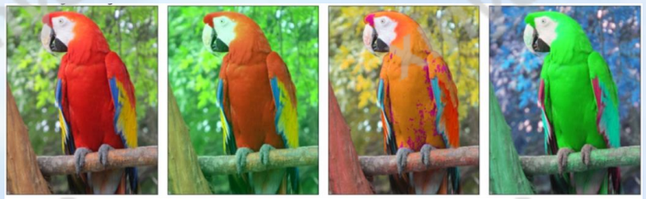
- 加入大量正常样本，类One-Class Classification；
- 针对样本偏差问题调整推理阈值。



One-Class Classification - SVDD

混杂对抗

- 设计大量数据增强方式；
- 将开源噪声集加入训练集。



Data Argument

信也科技杯 | 信也 FINVOLUTION

目录

1. 团队介绍
2. 赛题理解
3. 整体框架
4. 关键技术
5. 应用评估
6. 总结收获

2024第九届信也科技杯
全球AI算法大赛

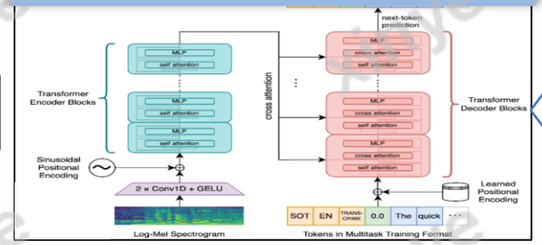
整体框架

开源数据微调模型

多样微调模型

OpenAI

Whisper

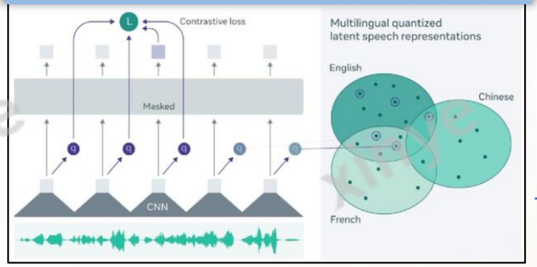


330M

多语言预训练模型

Meta

MMS



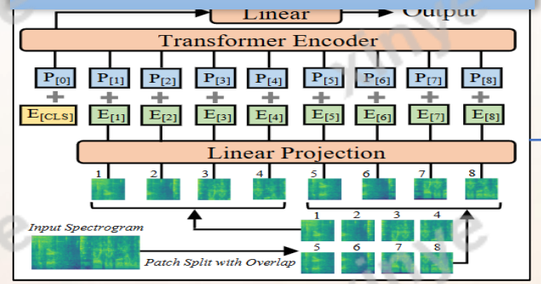
900M

400M

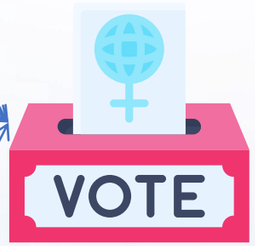
树模型

MIT

AST



110M



信也科技杯 | 信也 FINVOLUTION

目录

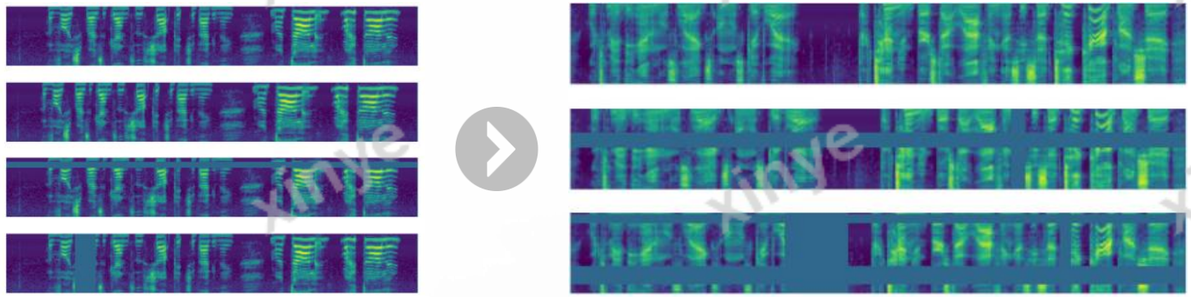
1. 团队介绍
2. 赛题理解
3. 整体框架
4. 关键技术
5. 应用评估
6. 总结收获

2024第九届信也科技杯
全球AI算法大赛

关键技术

数据增强

```
def set_model(model):
    model.config.apply_spec_augment = True
    model.config.mask_feature_prob = 0.1
    model.config.mask_time_prob = 0.1
    return model
```



时域频域增强示意图[1]

Frequency masking：选择一个从0到频率掩蔽参数F的均匀分布中的值f，然后掩蔽连续的f个梅尔频率通道$[f_0, f_0 + f)$，其中f_0从$[0, \nu - f)$中选择，ν是梅尔频率通道的数量。

Time masking：选择一个从0到时间掩蔽参数T的均匀分布中的值t，然后掩蔽连续的t个时间步$[t_0, t_0 + t)$，其中t_0从$[0, \tau - t)$中选择，τ是时间步的数量。

实践优势

遮蔽和扭曲操作迫使模型学习更加鲁棒的特征，这些特征在面对噪声和变形时依然有效。并且增加训练数据的多样性，扩展了模型看到的数据分布，使得模型在未见过的数据上表现更好。

[1] Park D S , Chan W , Zhang Y ,et al.SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition[J]. 2019.DOI:10.21437/Interspeech.2019-2680.

关键技术

为什么Mixup有效

提升了**0.02+**的F1

Mixup原理：Mixup通过在两个样本之间进行凸组合来创建新的虚拟样本。如果x_i和x_j是来自训练集的两个样本，λ是从指定分布中抽取的随机变量，则新的虚拟样本可以表示为$\tilde{x} = \lambda x_i + (1-\lambda)x_j$。

经验风险最小化(ERM)的问题：传统上，为了实现ERM就要对每一个样本降Loss，因此模型会尽力拟合尽可能多的样本，其实是“Memorize”，这和“Generalize”相对。

Mixup是对近邻风险进行优化：Mixup增强下的优化过程可以看作是对传统经验风险最小化ERM的一种泛化。实际为近邻风险最小化，损失函数期望可以写为：

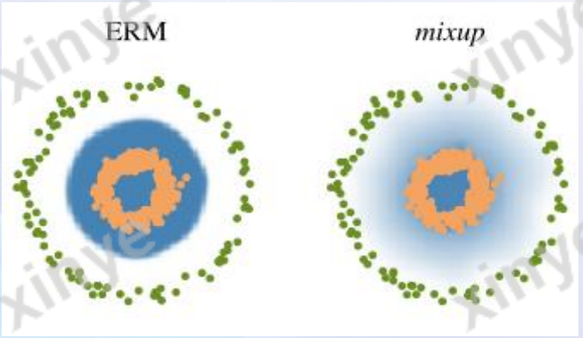
$$R_{mixup}(f) = E_{(x,y)\sim\mu}[l(f(\tilde{x}), \tilde{y})]$$

其中f是模型预测函数，l是损失函数，μ是Mixup生成的虚拟样本的分布。

对抗性样本的鲁棒性：对抗性样本是通过在合法样本上添加小的通过梯度上升得到扰动来生成的。Mixup通过鼓励模型在输入空间中平滑变化，倾向于在输入空间中有更小的梯度范数，这可以减少对抗扰动的影响。对于小的对抗性扰动δ，Mixup模型的鲁棒性可以表示为$f(\tilde{x}+\delta) \approx f(\tilde{x})$。

结论：Mixup其实就是告诉训练器，尽可能训练出线性的边界来。这样就会减少过拟合了。

```python
# 单样本数据增强
if np.random.rand() < 0.1:
    wave = clip(wave) # 裁剪
if np.random.rand() < 0.1:
    wave = echo(wave) # 回响
if np.random.rand() < 0.2:
    wave = add_noise(wave) # 加噪
if np.random.rand() < 0.2:
    wave = resample(wave) # 变频
if np.random.rand() < 0.05:
    wave, label, mixcnt = pitch_shift(wave), 0, 1 # 变调
if np.random.rand() < 0.05:
    wave, label, mixcnt = speedup(wave), 0, 1 # 变速
# mixup增强
if len(wave) <= 5 * 16000 and mixcnt == 0 and np.random.rand() < 0.3:
    mixcnt = np.random.randint(1,4)
    for _ in range(mixcnt):
        if len(wave) >= 15 * 16000:
            break
        if np.random.rand() < 0.8:
            pos = get_data(inputs, labels, mixups, 1)
            wave = mixup(wave, pos)
        else:
            neg = get_data(inputs, labels, mixups, 0)
            wave = mixup(wave, neg)
            label = 0
```



Mixup和ERM对比示意图[2]

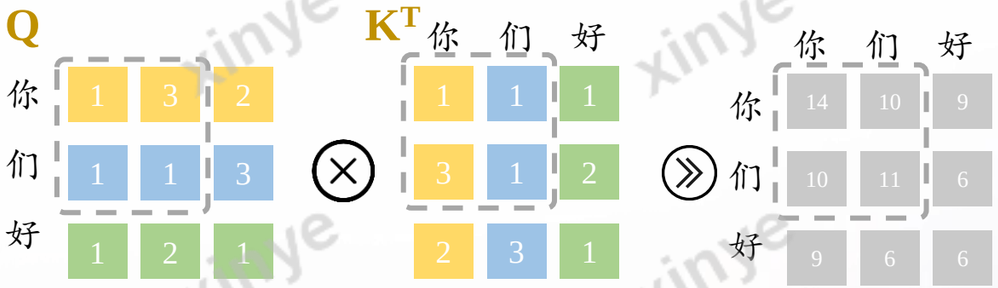
*绿色: Class 0. 橙色: Class 1. 蓝色:判定为0的区域.

[2] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

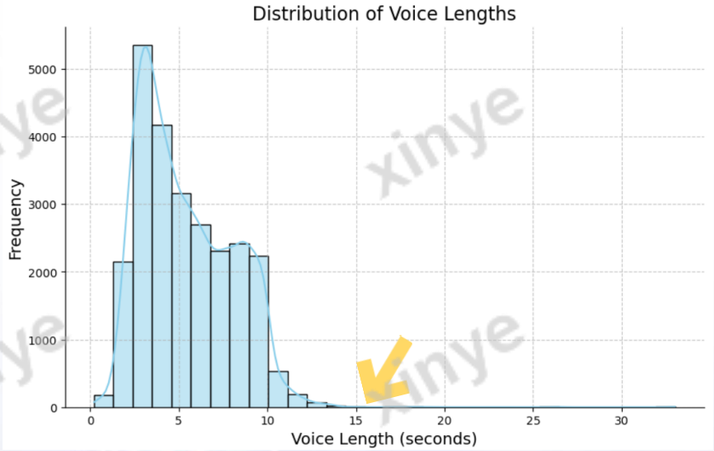
关键技术

注意力增强

提升了**0.01**的F1



self-attention[1]QK计算示意图
*计算复杂度为$n^2 \cdot d$



训练集语音时长频数分布图

理论优势

```
while len(wave) >= 15 * 16000:
    wave = clip(wave)
```

权重计算范围：当输入长度从30减少到15时，每个查询向量的注意力权重只需要考虑前15个时间步的键和值。这样，模型的计算复杂度从 $O(30^2 \cdot d_k)$ 降到 $O(15^2 \cdot d_k)$。其中 d_k 是键值的维度。

注意力分布变化：缩短输入序列长度会改变注意力分布的稀疏性。对于一个较长的输入序列（长度为30），注意力权重可能会分散在更多的时间步上。而对于较短的输入序列（长度为15），注意力权重会更加集中在有限的时间步内。这可以让模型更容易关注细节。

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

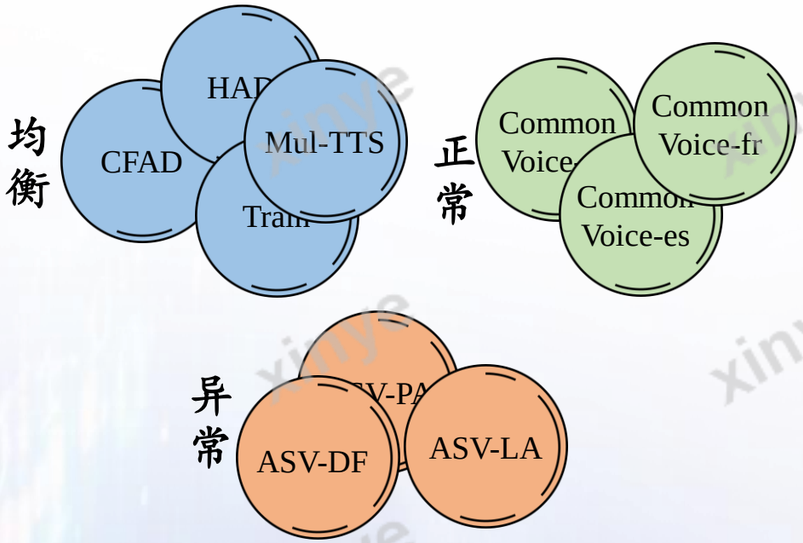
关键技术

模型融合

大量异常样本 » 严格 76.9

均衡样本 » 中庸 78.7 » 79.1 » 79.2

大量正常样本 » 谨慎 77.2



理论优势

假设：基模型的期望为 μ，方差σ^2，m个模型的权重为r，两两模型间的相关系数ρ相等。对于 Bagging 来说，每个基模型的权重等于 1/m 且期望近似相等，故我们可以得到Bagging模型F的总体方差：

$$Var(F) = \sigma^2(1-\rho)\cdot mr^2 + \sigma^2\rho\cdot mr^2 = \sigma^2(1-\rho)\cdot m/m^2 + \sigma^2\rho\cdot m/m^2 = \sigma^2(1-\rho)/m + \sigma^2\rho \le \sigma^2$$

结论：通过上式我们可以看到整体模型的方差小于等于基模型的方差，当且仅当ρ= 1 时取等号，随着基模型数量增多，整体模型的方差减少，从而防止过拟合的能力增强，模型的准确度得到提高。*Voting算法与Bagging算法类似。

信也科技杯 | 信也 FINVOLUTION

目录

1. 团队介绍
2. 赛题理解
3. 整体框架
4. 关键技术
5. 应用评估
6. 总结收获

2024第九届信也科技杯
全球AI算法大赛

应用评估

硬件评估

Model	Param	RAM	Speed	F1 Score	Finals Rank
Whisper-base-fp64	74M	0.3GB	32x	~0.760	7
Whisper-small-fp64	244M	1GB	12x	0.777	5
Whisper-large-fp64	1550M	10GB	2x	~0.771	6
Whisper-small-fp16	244M	0.3GB	18x	0.778	5
Whisper-small-attn-argument	244M	0.3GB	~32x	0.787	3
Whisper-small-Voting	-	0.9GB	6x	0.791	3
Voting Model(best)	-	2GB	1x	0.792	3

结论：为了获得最好的F1，我们采取了损失效率的复杂融合方法，但是我们使用的基础模型在保证准确性的同时有极快的推理效率。

信也科技杯 | 信也 FINVOLUTION

目录

1. 团队介绍
2. 赛题理解
3. 整体框架
4. 关键技术
5. 应用评估
6. 总结收获

2024第九届信也科技杯
全球AI算法大赛

总结收获

总结

- 分析：深入分析了音频的基础统计量，如音频长度，语言分布等。并针对分析设计了后续方案。
- 模型：使用三种多语言预训练模型，以whisper为主。最后使用voting算法对不同严格程度的模型进行融合，实现三个臭皮匠顶一个诸葛亮。
- 技巧：针对测试集的多种奇妙构造，设计了复杂的数据增强方式，旨在增加数据多样性并缓和分布偏移。
- 数据：选取大量多语言伪造数据以及真实数据，以及最新的伪造数据，旨在贴近测试集分布。

收获

- 感谢：信也科技作为互联网金融界巨头，打造了一个优秀的年度赛事”信也科技杯”，助力领域内青年人才成长，探索科技赋能的更多可能。感谢信也科技提供机会，让我们接触科技前沿，在课堂外参与非常有价值的实践。
- 学习：我们第一次打audio系列的比赛，近期跟管理员同学、赛事负责老师学习了很多新知识，收获颇多。尤其感谢biu老师在答疑群里悉心解答各位选手的问题。
- 合作：团队成员通力合作，度过了各种难关。每个成员在自己的小领域中，都出色的完成了分析和建模任务，最终搭成框架，满意的完成了工作。

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