TOP1 主办方 msup

声音技术的未来
大模型带来的音频算法革新

张俊博

小米AI实验室 语音技术专家

讲师简介

"

小米语音技术专家。

博士毕业于中国科学院声学研究所，多年从事智能语音技术的研究和应用，在语音识别、发音评测、语音合成、音频标记等领域都做过深入的工作，在顶级会议和期刊发表论文 30 余篇，著有出版物《Kaldi 语音识别实战》。

目前在小米负责若干项声学语音新技术的研发。

"

内容提要

- 对大模型的思考
- 小米的音频大模型探索

对大模型的思考

是“发现”，而不是“发明”

```
LlamaModel(
  (embed_tokens): Embedding(32000, 256)
  (layers): ModuleList(
    (0-7): 8 x LlamaDecoderLayer(
      (self_attn): LlamaAttention(
        (q_proj): Linear(in_features=256, out_features=256, bias=False)
        (k_proj): Linear(in_features=256, out_features=1024, bias=False)
        (v_proj): Linear(in_features=256, out_features=1024, bias=False)
        (o_proj): Linear(in_features=256, out_features=256, bias=False)
        (rotary_emb): LlamaRotaryEmbedding()
      )
      (mlp): LlamaMLP(
        (gate_proj): Linear(in_features=256, out_features=11008, bias=False)
        (up_proj): Linear(in_features=256, out_features=11008, bias=False)
        (down_proj): Linear(in_features=11008, out_features=256, bias=False)
        (act_fn): SiLUActivation()
      )
      (input_layernorm): LlamaRMSNorm()
      (post_attention_layernorm): LlamaRMSNorm()
    )
  )
  (norm): LlamaRMSNorm()
)
```

Llama2 模型：没有任何模型结构上的创新

原理上是量变，效果上是质变

无法解释，只好说“涌现”

大模型的成功，证明了这样的路线是可行的

为 AI 研究指明了方向

为什么大模型具备如此神奇的能力?

不知道

人类对它的原理还远远称不上理解

但大模型研发并没有技术原理上的门槛

虽然不知道麦克斯韦方程组

不妨碍古人发明指南针

虽然暂时未能全面理解大模型

不妨碍我们做出更强的大模型

雨后春笋般的大模型研发

图片来源 SuperCLUE

小米自研大语言模型

本地化、轻量部署

	平均分	STEM	人文学科	社会科学	其他	中国特定主题
MiLM-1.3B	**50.79**	**40.51**	**54.82**	**54.15**	**53.99**	**52.26**
Baichuan-13B	54.63	42.04	60.49	59.55	56.6	55.72
ChatGLM 2-6B	49.95	41.28	52.85	53.37	52.24	50.58

CMMLU 中文多任务语音理解评估，2023年8月10日数据

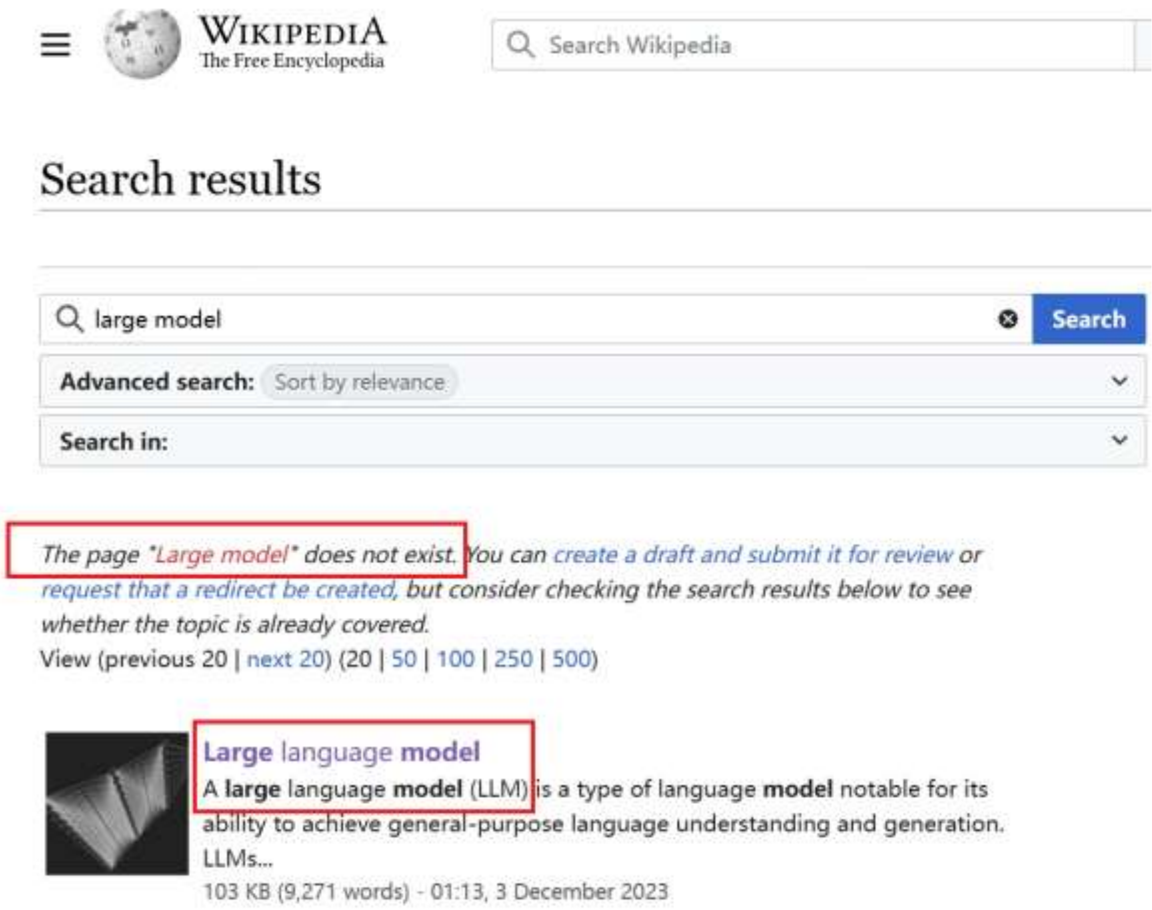
手机端侧大模型部分场景媲美云端

大模型 == 大语言模型？

文本形式训练数据相对更易获取和处理

大模型首先以文本模态出现

但人类更倾向于使用视觉和声音交互



GPT4-V(ision)

OpenAI

User

What is unusual about this image?

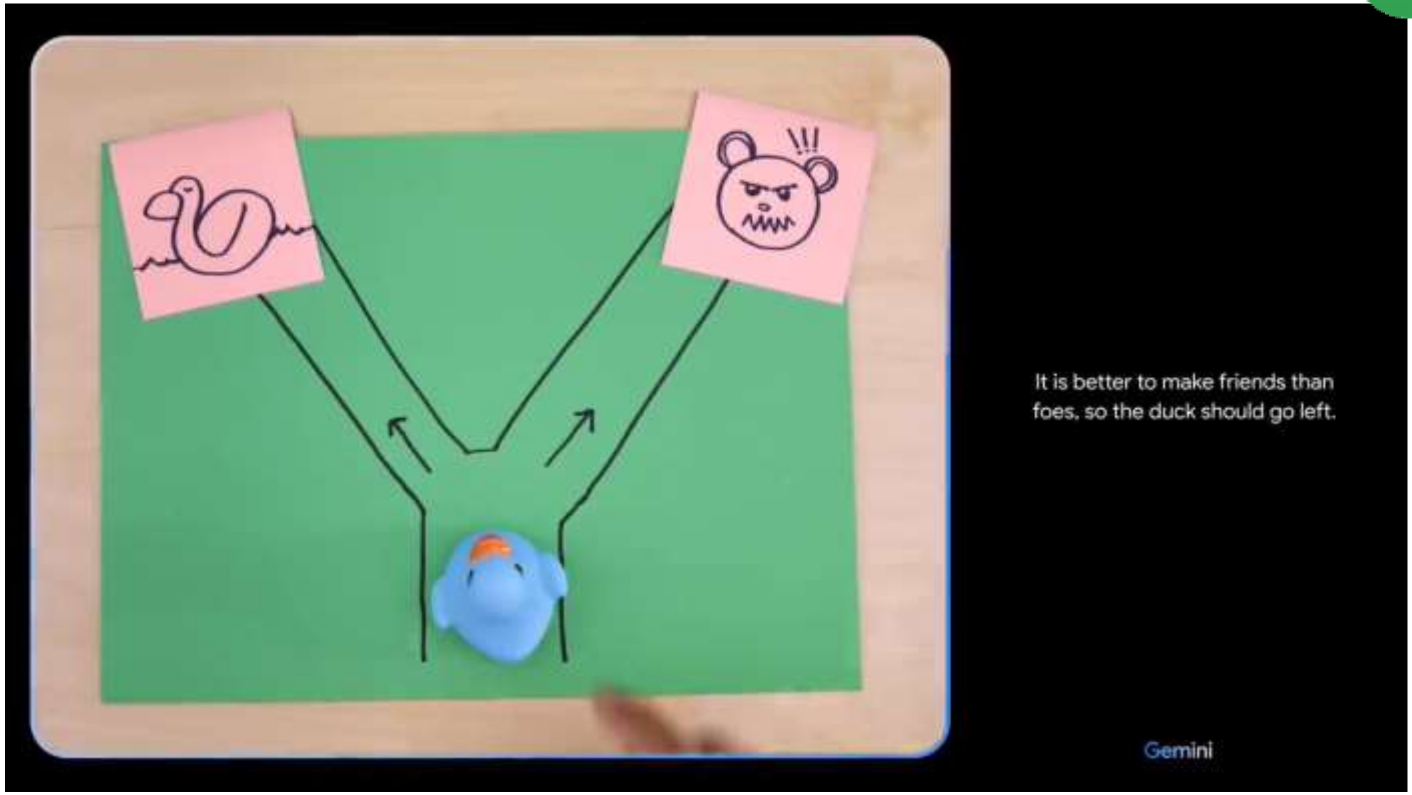
GPT-4

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Gemini

令人震惊的多模态能力

强人工智能已实现?



小米的音频大模型探索

AI 时代的小米

视觉 | 声学语音 | NLP | 知识图谱 | 机器学习 | 大模型 | 多模态

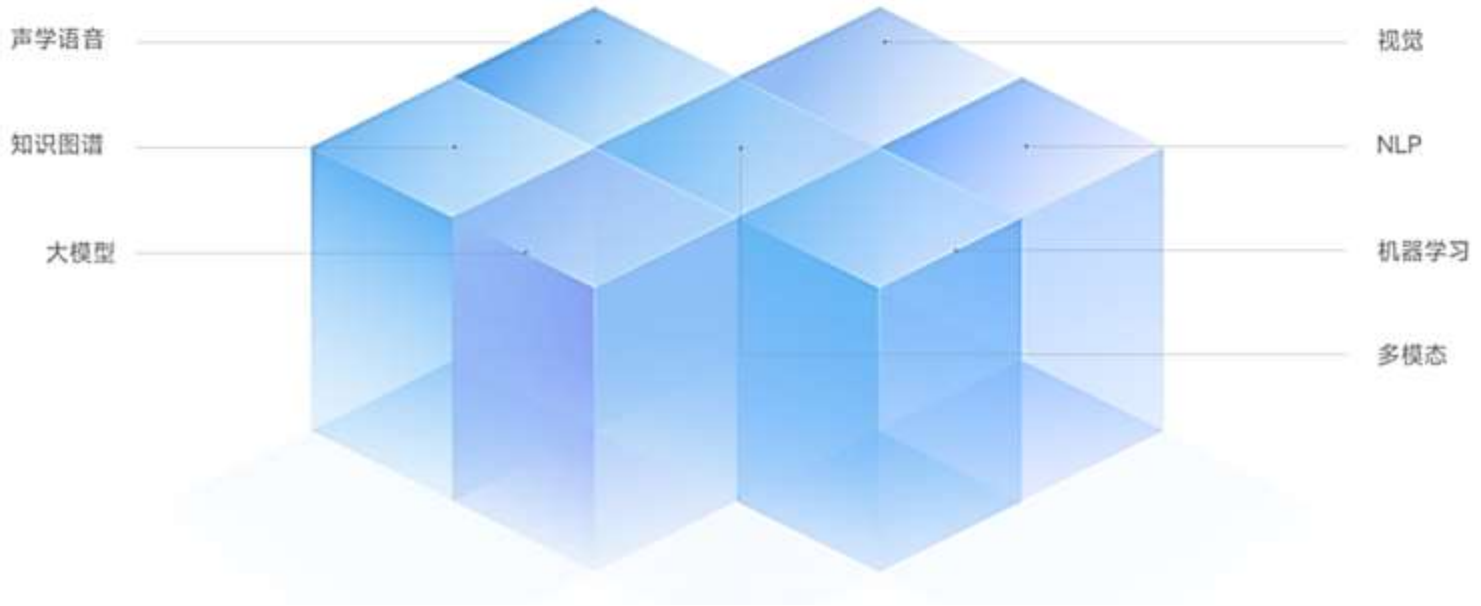
全球最大消费级 IoT 平台

6.99 亿

IoT 平台已连接设备数

1370 万

拥有 5 个及以上小米 IoT 设备的用户数



小米声学语音技术

理解

语音
- 语音识别
- 命令词唤醒
- 声纹识别

声音
- 声音事件检测
- 声音场景分析
- 音频描述

生成

语音
- 语音合成
- 说话人转换
- 歌声合成

声音
- 声音成分编辑
- 音效生成
- 音乐创作

小爱同学背后的语音识别技术框架

在手机和 IoT 设备上针对垂域的识别率极高，且已经探索出成熟的迭代优化流程

但是！这不是大模型时代的方案

技术革新势在必行

Whisper：大模型语音识别

OpenAI

Whisper examples: Speed talking

This is the Micro Machine Man presenting the most midget miniature motorcade of Micro Machines. Each one has dramatic details, terrific trim, precision paint jobs, plus incredible Micro Machine Pocket Play Sets. There's a police station, fire station, restaurant, service station, and more. Perfect pocket portables to take any place. And there are many miniature play sets to play with, and each one comes with its own special edition Micro Machine vehicle and fun, fantastic features that miraculously move. Raise the boatlift at the airport marina. Man the gun turret at the army base. Clean your car at the car wash. Raise the toll bridge. And these play sets fit together to form a Micro Machine world. Micro Machine Pocket Play Sets, so tremendously tiny, so perfectly precise, so dazzlingly detailed, you'll want to pocket them all. Micro Machines are Micro Machine Pocket Play Sets sold separately from Galoob. The smaller they are, the better they are.

Whisper examples: French

Whisper is an automatic speech recognition system based on 680,000 hours of multilingual and multitasking data collected on the Internet. We establish that the use of such a number of data is such a diversity and the reason why our system is able to understand many accents, regardless of the background noise, to understand technical vocabulary and to successfully translate from various languages into English. We distribute as a free software the source code for our models and for the inference, so that it can serve as a starting point to build useful applications and to help progress research in speech processing.

Whisper examples: K-Pop

While darkness was my everything
I ran so hard that I ran out of breath
Never say time's up
Like the end of the boundary
Because my end is not the end

Whisper examples: Accent

One of the most famous landmarks on the Borders, it's three hills and the myth is that Merlin, the magician, split one hill into three and left the two hills at the back of us which you can see. The weather's never good though, we stay on the Borders with the mists on the Yildens, we never get the good weather and as you can see today there's no sunshine, it's a typical Scottish Borders day.

Whisper 原理有何不同?

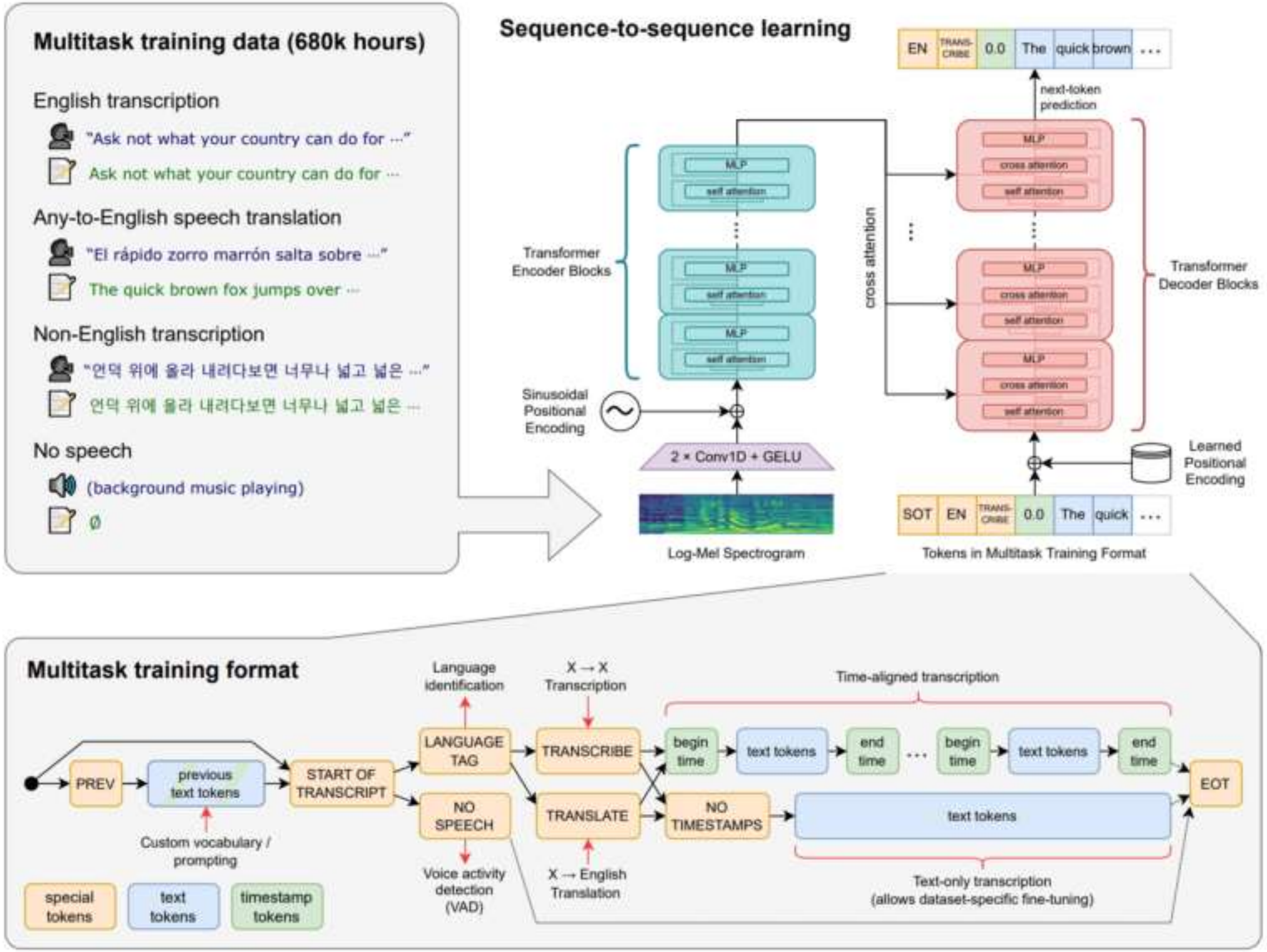
更先进的模型结构？ No

模型结构并无不同

多语种训练数据

带有多任务标签

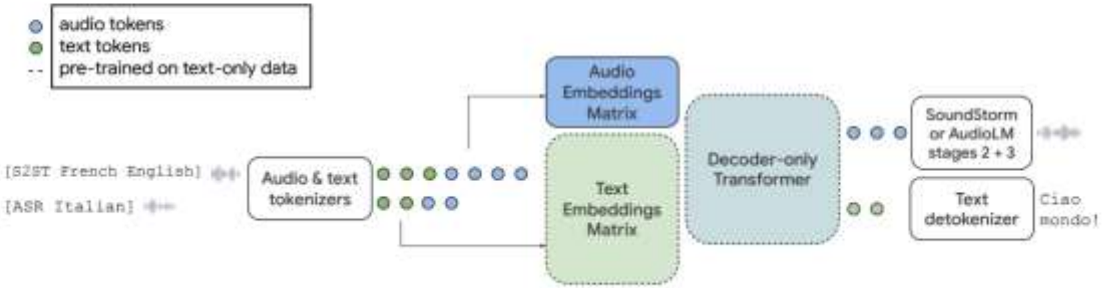
680,000 hours of multilingual and multitask supervised data collected from the web



AudioPaLM：多语种语音直译

大语言模型作为模型骨架和初始化参数

多语种音频和文本数据迭代训练



Mandarin / 中文

AudioPaLM

Original 我非常开心和你合作

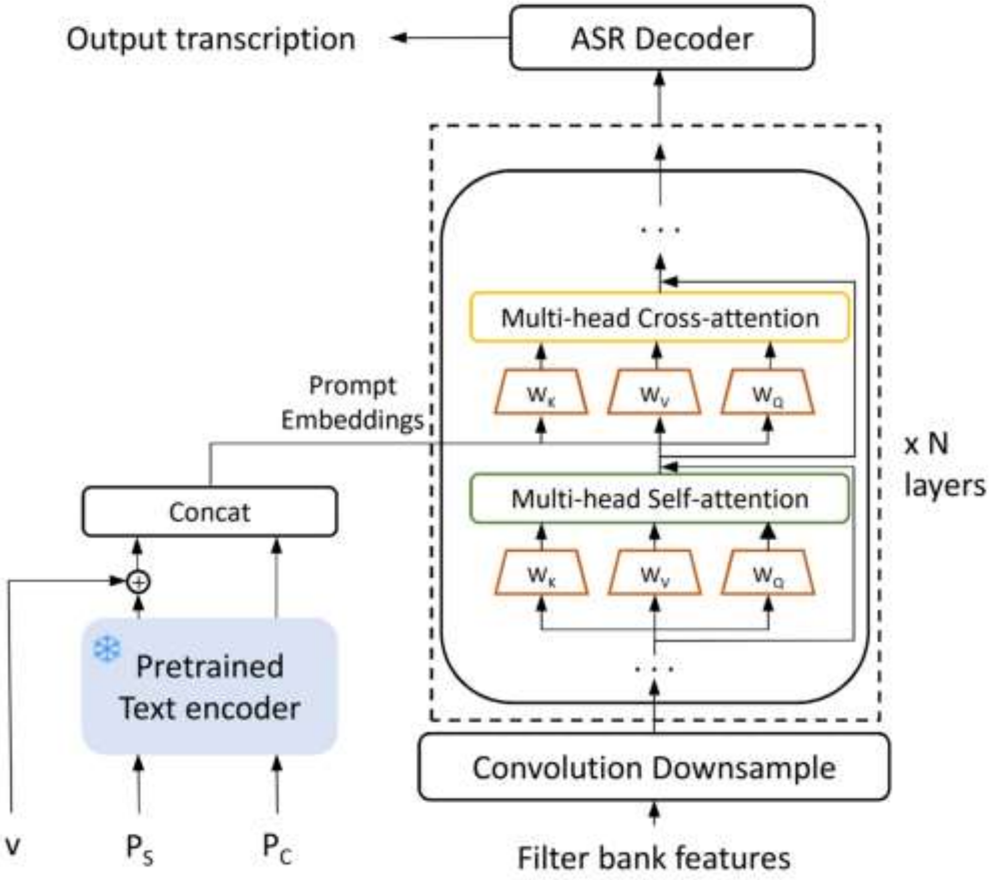
Translation with AudioPaLM I am very happy to work with you

小米 Prompt-ASR

用 prompt 约束语音识别领域，提升识别率

把大语言模型输出通过 cross-attention 联入 encoder

Style Prompt	WITHOUT CASING OR PUNCTUATION
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	TODAY'S MATCH IS BETWEEN REAL MADRID AND LIVERPOOL
Style Prompt	Mixed-cased English with punctuation
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	Today's match is between Real Madrid and Liverpool.



基于大模型的语音合成

更加自然

视频来源
https://www.bilibili.com/video/BV1e84y1U7j4

支持 Prompt 定制

Microsoft VALL-E X

Look a little closer while our guide lets the light of his lamp fall upon the black wall at your side.

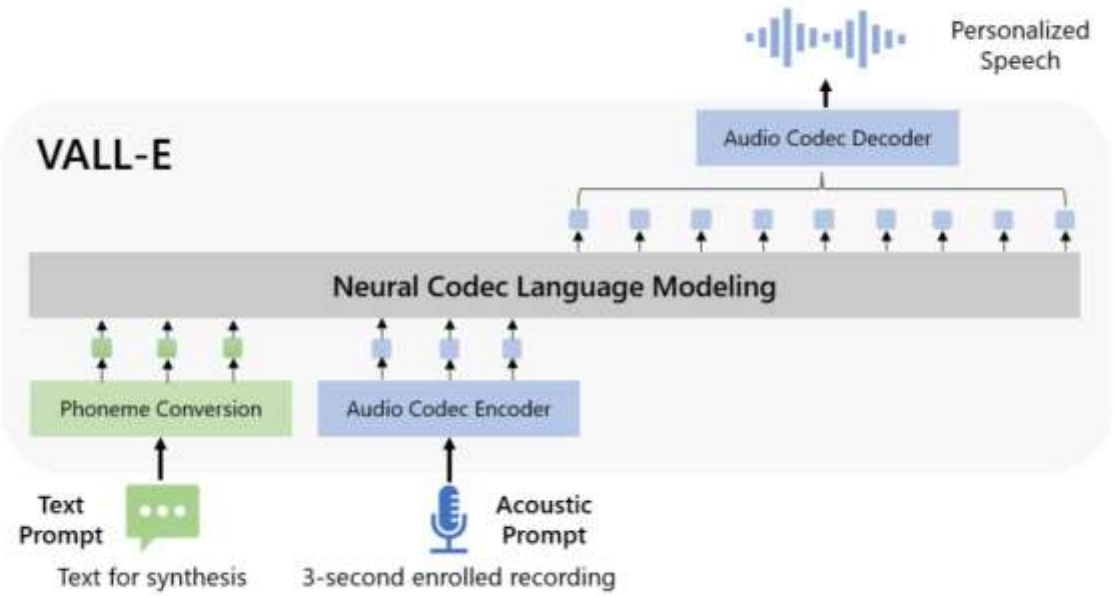
baseline

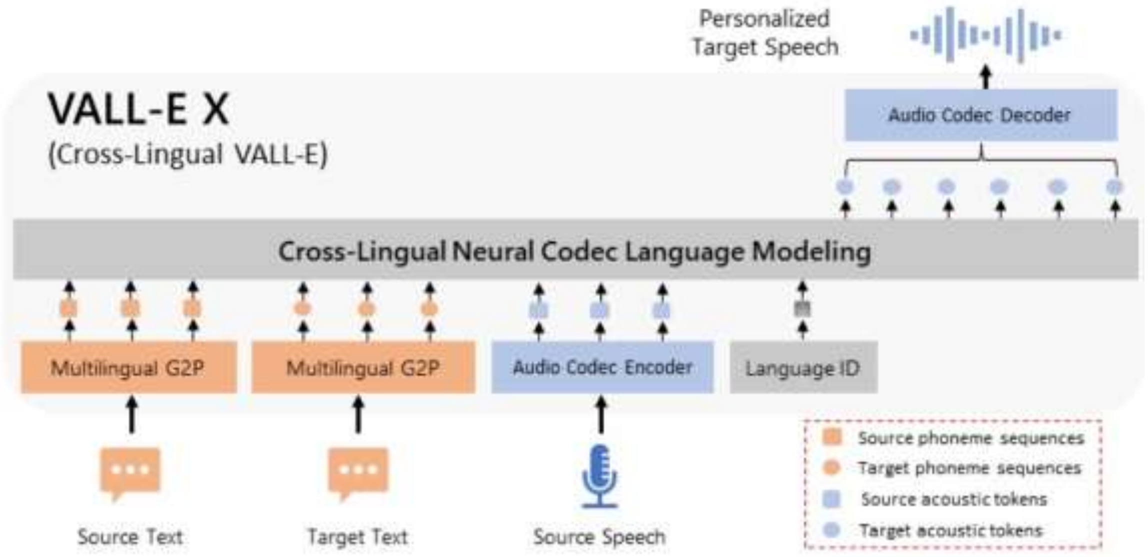
中文说话人

合成效果

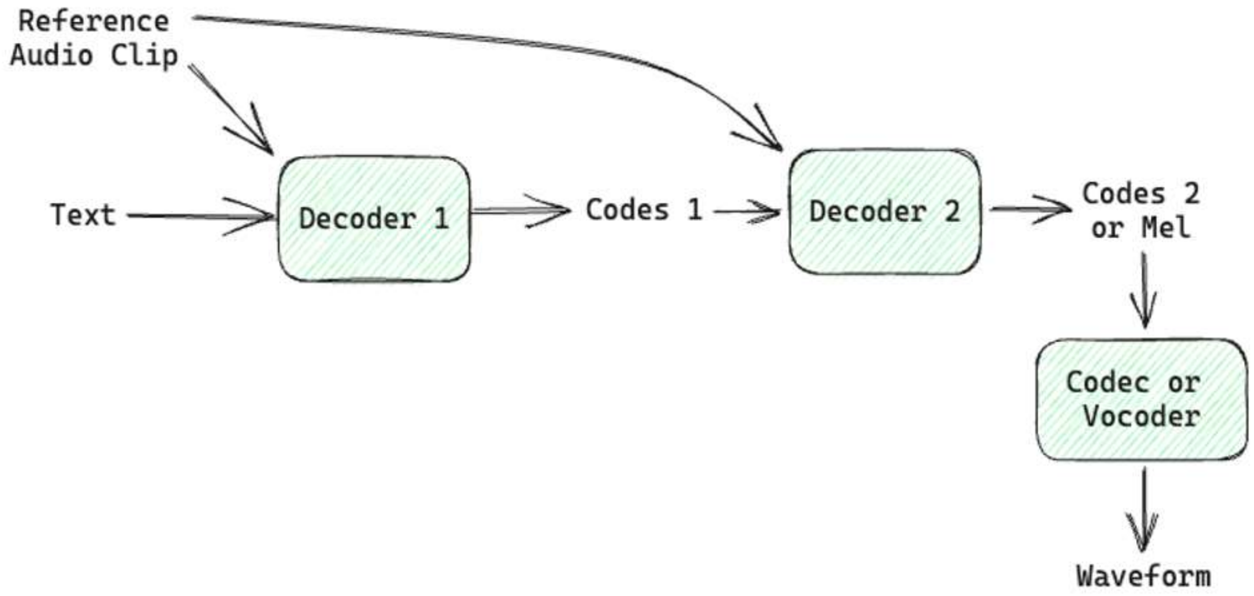
VALL-E (X) 算法框架

Microsoft





基于大模型的小米自然语音 TTS

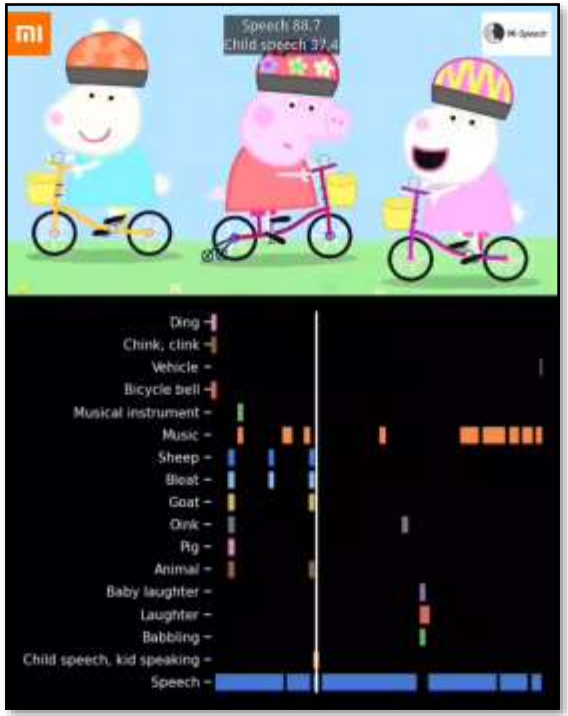


使用小爱默认音色

说话人迁移

(prompt)

小米声音识别技术



目前支持 85 种声音事件

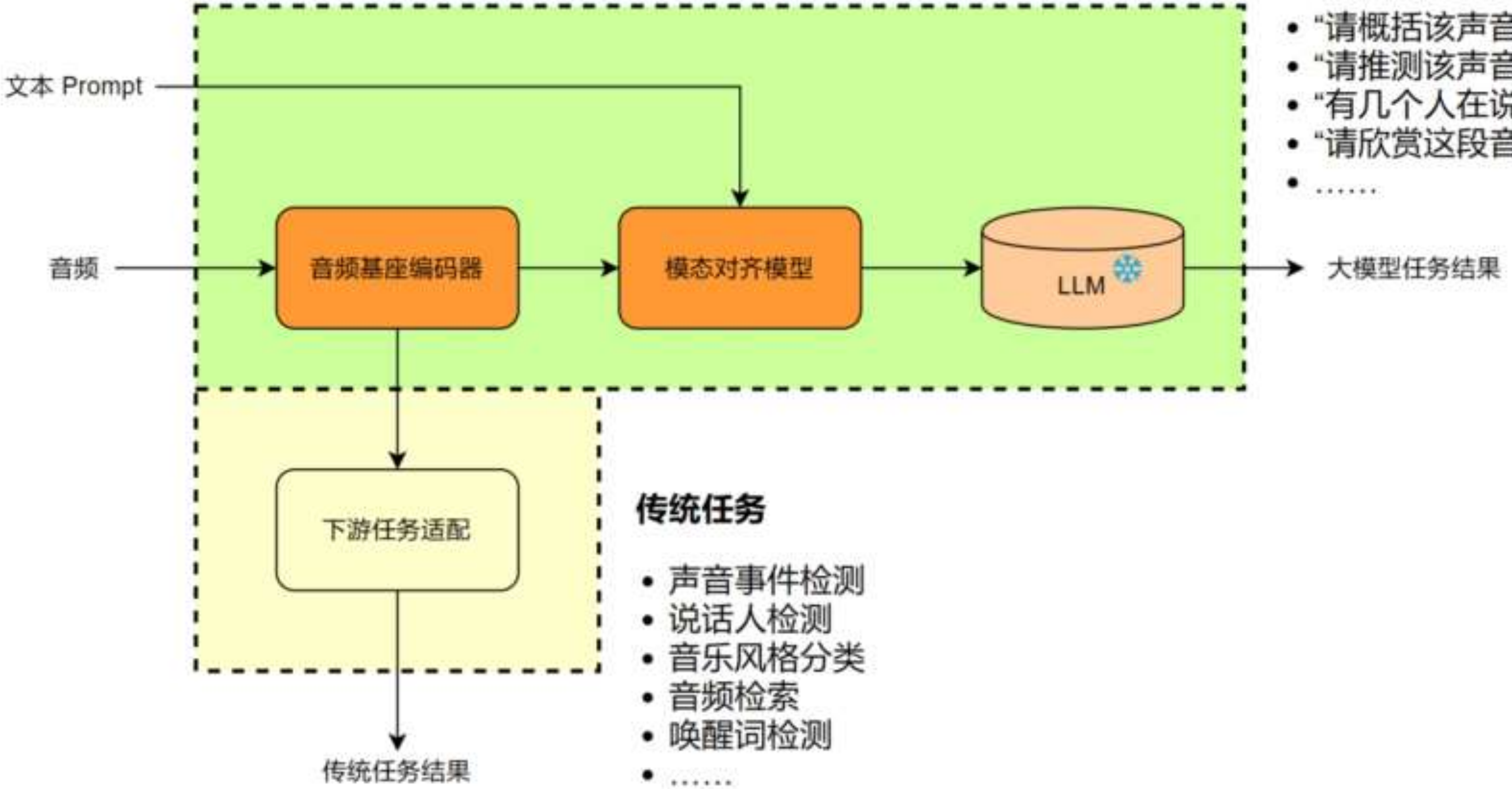
大模型时代的声音理解

请描述这段声音的主要内容?

他们在讨论什么问题?结论是什么?

现场的气氛怎样?

我们的算法框架



大模型任务

- “请概括该声音的内容？”
- “请推测该声音发生的地点？”
- “有几个人在说话、他们在争论什么问题、情绪如何？”
- “请欣赏这段音乐，并针对其艺术水平写一段300字的评论。”
- ……

传统任务

- 声音事件检测
- 说话人检测
- 音乐风格分类
- 音频检索
- 唤醒词检测
- ……

基座音频编码器

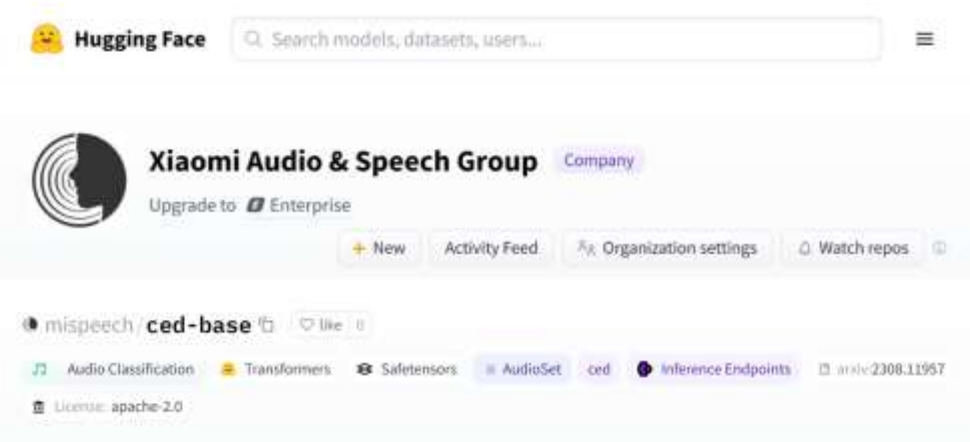
训练数据时长超过30年

参数量超过10亿

正在探索百亿参数量的模型

独创的一致性集成蒸馏技术

论文已被 ICASSP 2024 接收



模型开源可下载

arXiv:2308.11957v2 [cs.SD] 8 Sep 2023

CED: CONSISTENT ENSEMBLE DISTILLATION FOR AUDIO TAGGING

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, Yujun Wang

Xiaomi Corporation, Beijing, China

ABSTRACT

Augmentation and knowledge distillation (KD) are well-established techniques employed in audio classification tasks, aimed at enhancing performance and reducing model sizes on the widely recognized Audioset (AS) benchmark. Although both techniques are effective individually, their combined use, called consistent teaching, hasn't been explored before. This paper proposes CED, a simple training framework that distils student models from large teacher ensembles with consistent teaching. To achieve this, CED efficiently stores logits as well as the augmentation methods on disk, making it scalable to large-scale datasets. Central to CED's efficacy is its label-free nature, meaning that only the stored logits are used for the optimization of a student model only requiring 0.3% additional disk space for AS. The study trains various transformer-based models, including a 10M parameter model achieving a 49.0 mean average precision (mAP) on AS. Pretrained models and code are available online.

Index Terms— audio tagging, audio classification, efficient data storage, teacher-student, knowledge distillation

1. INTRODUCTION

小米声音识别算法
性能在国际上
排名第一

模型名称	机构	mAP	参数量（M）
CED	小米	50.0	86
CED-mini	小米	49.0	10
Efficient-AT	JKU	48.7	68
BEATs	微软	48.6	90
Audio-MAE	Meta、CMU	47.3	86
ConvNeXt	ANITI 等	47.1	28
HTS-AT	UCSD、字节跳动	47.1	31
MaskSpec	[illegible]	47.1	86
AST	MIT	45.9	86
PANNs-CNN14	萨里大学、字节跳动	45.1	81
PSL-Mbv2	小米	40.3	3

*2023年9月12日结果

基座音频编码器的多任务应用

HEAR BENCHMARK

Holistic Evaluation of Audio Representations

HEAR Leaderboard

Model	URL	Submission Date	Beehive	Beijing Opera	CREMA-D	DCASE 2016	ESC-50	FSD50K	GTZAN Genre	GTZAN Music/Speech
RedRice eed_base		2023-09-26	0.4835	0.9660	0.6910	0.9219	0.9665	0.6548	0.8900	0.9436
GURA Fuse Cat H+w+C		HEAR 2021		0.9660	0.7474	0.8260	0.7335	0.4197	0.8050	0.9282
RedRice eed_small		2023-09-26	0.5170	0.9660	0.6664	0.9163	0.9595	0.6433	0.8950	0.9122
GURA Fuse Cat H+w+C (time)		HEAR 2021		0.9618	0.7427	0.8260	0.6535	0.3742	0.7600	0.9436
RedRice eed_mini		2023-09-26	0.5017	0.9618	0.6526	0.9066	0.9535	0.6388	0.9030	0.9449
Logitech AI SERAB BYOL-S		HEAR 2021	0.5487	0.9533	0.6560	0.6421	0.8050	0.5088	0.8370	0.9385
CP-JKU mn40_as (s1+l+all_se)		2023-01-05	0.5027	0.9534	0.6400	0.8130	0.9615	0.6312	0.9080	0.9840
GURA Fuse Hubert		HEAR 2021		0.9493	0.7521	0.8259	0.7435	0.4132	0.7960	0.9350
GURA Cat H+w+C		HEAR 2021		0.9363	0.6394	0.6808	0.5110	0.3144	0.7220	0.9609
OpenL3		HEAR 2021	0.6040	0.9746	0.5497	0.8328	0.7505	0.4470	0.8790	0.9686
GURA Fuse wav2vec2		HEAR 2021		0.9449	0.6924	0.7983	0.6050	0.4028	0.7930	0.9532
GURA Avg H+w+C		HEAR 2021		0.9448	0.5473	0.6236	0.4495	0.2636	0.7060	0.9365
CP-JKU PaSST 2lvl+mel		HEAR 2021		0.9660	0.6104	0.9254	0.9475	0.6409	0.8830	0.9769
CP-JKU PaSST 2lvl		HEAR 2021		0.9660	0.6104	0.9132	0.9475	0.6409	0.8830	0.9760
GURA Avg Hubert+Crepe		HEAR 2021		0.9323	0.5397	0.6104	0.4365	0.2528	0.6980	0.9455
RedRice EfficientNet-B2		HEAR 2021	0.5332	0.9535	0.5746	0.7901	0.9345	0.6071	0.8780	0.9679
CP-JKU PaSST base		HEAR 2021		0.9660	0.6104	0.7879	0.9475	0.6409	0.8830	0.9769
GURA Cat wav2vec2+crepe		HEAR 2021		0.9198	0.4598	0.5855	0.3430	0.2341	0.6810	0.9378
wav2vec2		HEAR 2021		0.9067	0.6562	0.6630	0.5610	0.3417	0.7800	0.9462

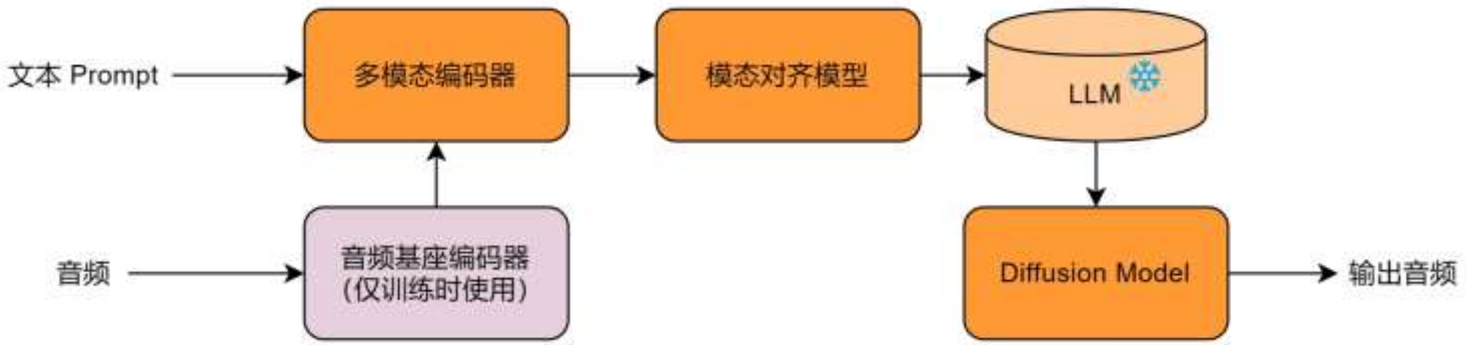
Model	Gunshot	Libricount	Maestro 5h	Mridangam Stroke	Mridangam Tonic	NSynth Pitch 50h	NSynth Pitch 5h	Speech commands 5h	Speech commands full	Vocal Imitation	VoxLingua107 top 10
RedRice eed_base	0.8929	0.6785	0.1476	0.9743	0.9655	0.8281	0.6820	0.8693	0.8967	0.2269	0.3857
GURA Fuse Cat H+w+C	0.9345	0.6972	0.4413	0.9725	0.9235	0.8852	0.8460	0.9605	0.9677	0.1969	0.7202
RedRice eed_small	0.9345	0.6559	0.1096	0.9682	0.9304	0.7995	0.6020	0.8092	0.8519	0.2192	0.3653
GURA Fuse Cat H+w+C (time)	0.9048	0.6591	0.4413	0.9753	0.9243	0.8909	0.8540	0.9509	0.9681	0.2152	0.6286
RedRice eed_mini	0.8601	0.6402	0.08289	0.9656	0.9132	0.7520	0.5560	0.7738	0.8196	0.2037	0.3467
Logitech AI SERAB BYOL-S	0.8571	0.7853	0.007860	0.9726	0.9285	0.7116	0.3960	0.9137	0.9481	0.1598	0.4578
CP-JKU mn40_as (s1+l+all_se)	0.9107	0.7253	0.01862	0.9713	0.9047	0.6355	0.3040	0.7767	0.8472	0.2019	0.3179
GURA Fuse Hubert	0.9286	0.6834	0.1657	0.9738	0.9090	0.6982	0.3820	0.9468	0.9571	0.1848	0.7140
GURA Cat H+w+C	0.8810	0.6390	0.4691	0.9378	0.8591	0.8968	0.8600	0.9270	0.9429	0.1114	0.4399
OpenL3	0.9494	0.6414	0.01650	0.9666	0.9369	0.7310	0.5600	0.6796	0.7634	0.07812	0.3313
GURA Fuse wav2vec2	0.9673	0.6526	0.1110	0.9623	0.8376	0.6059	0.3300	0.9571	0.9687	0.1740	0.7058
GURA Avg H+w+C	0.8571	0.6309	0.4709	0.9144	0.8373	0.8963	0.8620	0.8225	0.8808	0.08415	0.3210
CP-JKU PaSST 2lvl+mel	0.9405	0.6601		0.9650	0.8194	0.5409	0.2560	0.6810	0.6387	0.1820	0.2593
CP-JKU PaSST 2lvl	0.9405	0.6601		0.9650	0.8194	0.5409	0.2560	0.6810	0.6387	0.1820	0.2593
GURA Avg Hubert+Crepe	0.8452	0.6224	0.4624	0.9170	0.8309	0.8966	0.8780	0.7372	0.8233	0.07895	0.2932
RedRice EfficientNet-B2	0.8780	0.6509	0.0001910	0.9485	0.8432	0.3914	0.1680	0.5734	0.6757	0.1385	0.2551
CP-JKU PaSST base	0.9405	0.6601		0.9650	0.8194	0.5409	0.2560	0.6810	0.6387	0.1820	0.2593
GURA Cat wav2vec2+crepe	0.8333	0.5694	0.4626	0.8977	0.8226	0.8990	0.8680	0.8853	0.9168	0.07640	0.3097
wav2vec2	0.8482	0.6921	0.03290	0.9432	0.8283	0.6530	0.4020	0.8382	0.8795	0.09096	0.4929

声音增强/编辑/生成

已有成果其实已经具备了部分大模型的能力

需要进一步整合

基于 Prompt 的声音生成



结语

- 大模型的成功为 AI 研究指明了方向
- 多任务统一学习可以带来真正的理解能力和强大的任务自推广能力
- 各任务的统一、各模态的统一是大势所趋

TOP1

主办方 msup

微信官方公众号：壹佰案例
关注查看更多年度实践案例