Microsoft Carnegie Mellon University Language Technologies Institute

NEURAL INFORMATION PROCESSING SYSTEMS

🐧 Pengi: Audio Language Model for Audio Tasks

Soham Deshmukh
Microsoft, CMU

Benjamin Elizalde
Microsoft

Rita Singh
CMU

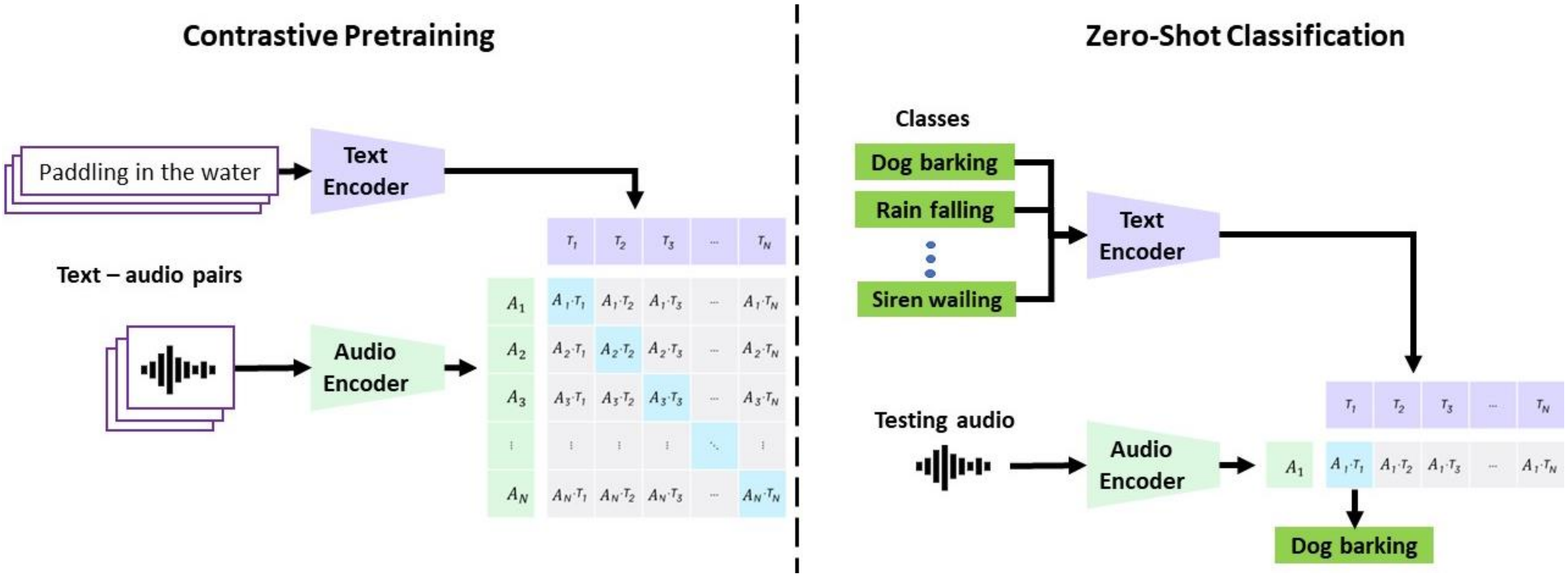
Huaming Wang
Microsoft

https://github.com/microsoft/pengi

Motivation

- **Contrastive Audio-Language models are used for zero-shot close-ended tasks, such as classification and retrieval**
- However, these models inherently lack the capacity to produce the requisite language for open-ended tasks, such as Audio Captioning or Audio Question & Answering
- Can we have a unified model that performs close-ended and open-ended tasks?

Motivation


Contrastive Pretraining
Zero-Shot Classification
Paddling in the water
Text Encoder
Text – audio pairs
Audio Encoder
T_1 T_2 T_3 ... T_N
A_1 $A_1 \cdot T_1$ $A_1 \cdot T_2$ $A_1 \cdot T_3$ ... $A_1 \cdot T_N$
A_2 $A_2 \cdot T_1$ $A_2 \cdot T_2$ $A_2 \cdot T_3$ ... $A_2 \cdot T_N$
A_3 $A_3 \cdot T_1$ $A_3 \cdot T_2$ $A_3 \cdot T_3$ ... $A_3 \cdot T_N$
$\vdots$ $\vdots$ $\vdots$ $\vdots$ $\ddots$ $\vdots$
A_N $A_N \cdot T_1$ $A_N \cdot T_2$ $A_N \cdot T_3$ ... $A_N \cdot T_N$
Classes
Dog barking
Rain falling
Siren wailing
Text Encoder
Testing audio
Audio Encoder
T_1 T_2 T_3 ... T_N
A_1 $A_1 \cdot T_1$ $A_1 \cdot T_2$ $A_1 \cdot T_3$ ... $A_1 \cdot T_N$
Dog barking

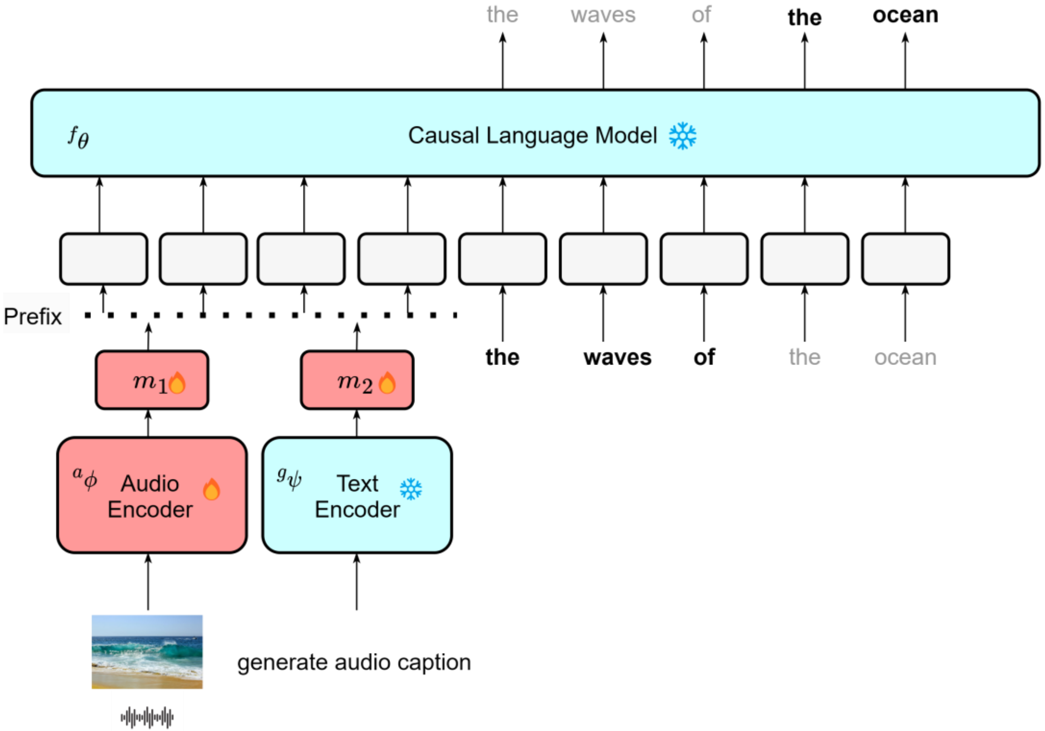
Motivation

- Contrastive Audio-Language models are used for zero-shot close-ended tasks, such as classification and retrieval
- **However, these models inherently lack the capacity to produce the requisite language for open-ended tasks, such as Audio Captioning or Audio Question Answering**
- Can we have a unified model that performs close-ended and open-ended tasks?

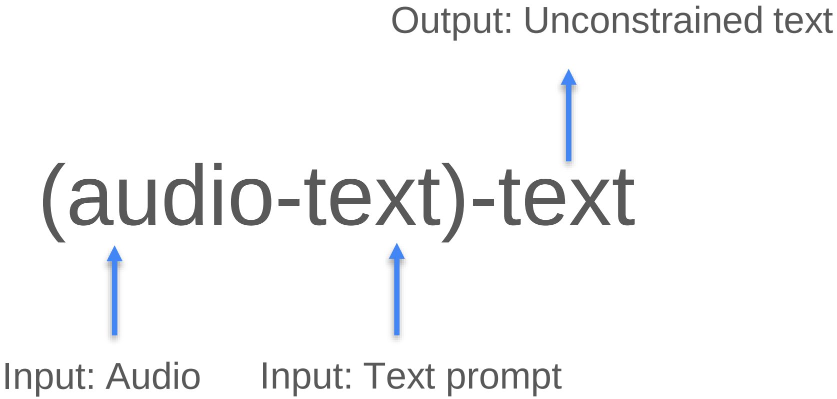
Motivation

- Contrastive Audio-Language models are used for zero-shot close-ended tasks, such as classification and retrieval
- However, these models inherently lack the capacity to produce the requisite language for open-ended tasks, such as Audio Captioning or Audio Question & Answering
- **Can we have a unified model that performs close-ended and open-ended tasks?**

Audio Language Model



Frame audio tasks as audio-text to text tasks



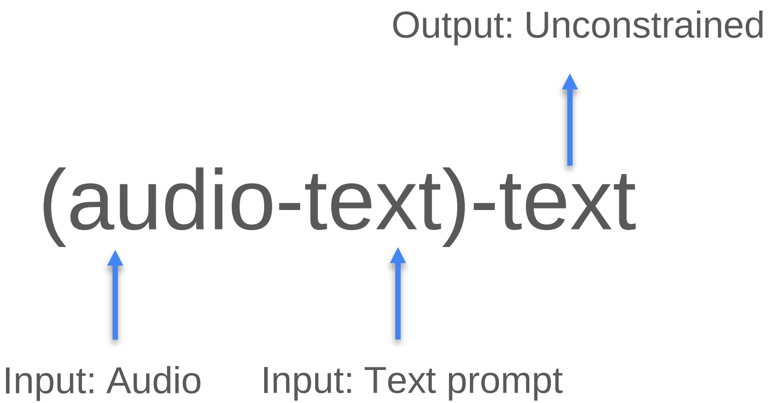
Output: Unconstrained text

(audio-text)-text

Input: Audio

Input: Text prompt

Audio-task templates for training



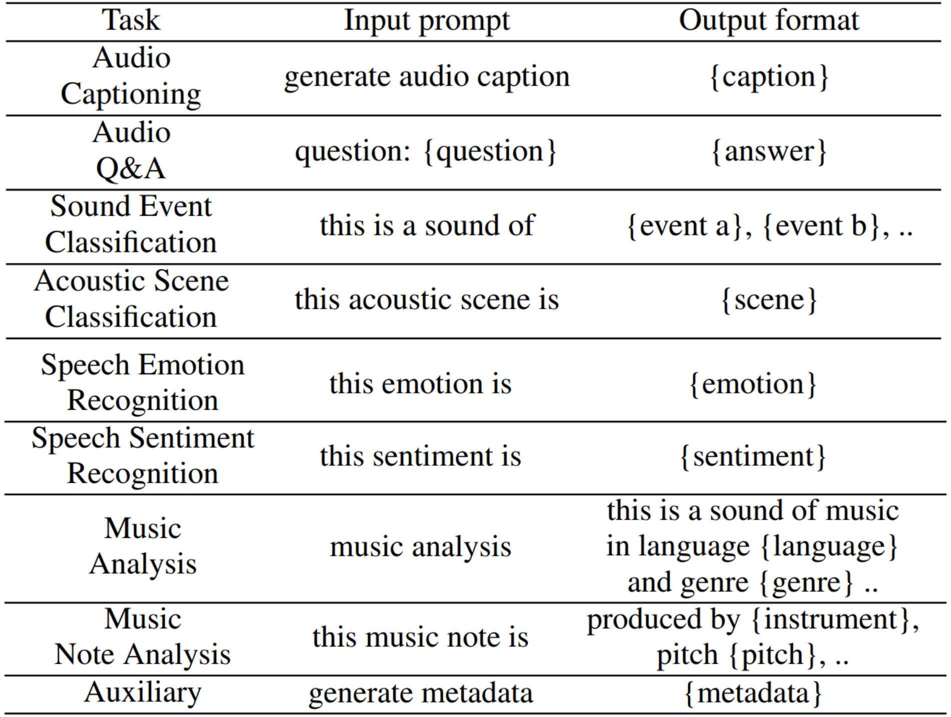
Output: Unconstrained

(audio-text)-text

Input: Audio

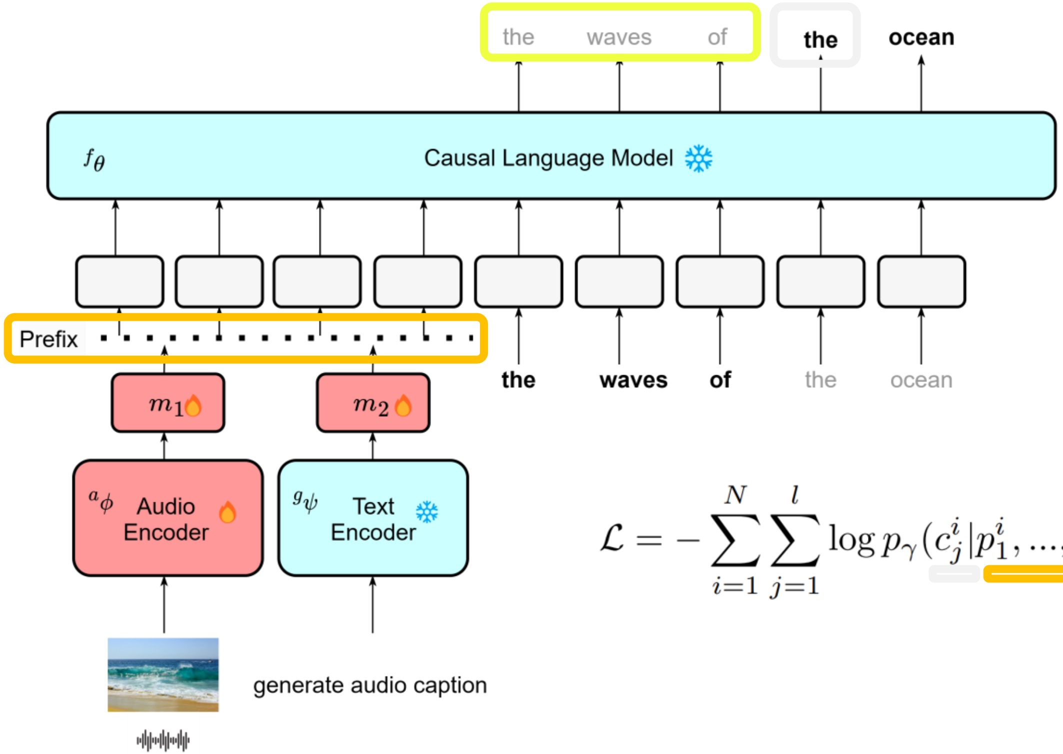
Input: Text prompt

3.4M (audio-text)-text triplets



Task	Input prompt	Output format
Audio Captioning	generate audio caption	{caption}
Audio Q&A	question: {question}	{answer}
Sound Event Classification	this is a sound of	{event a}, {event b}, ..
Acoustic Scene Classification	this acoustic scene is	{scene}
Speech Emotion Recognition	this emotion is	{emotion}
Speech Sentiment Recognition	this sentiment is	{sentiment}
Music Analysis	music analysis	this is a sound of music in language {language} and genre {genre} ..
Music Note Analysis	this music note is	produced by {instrument}, pitch {pitch}, ..
Auxiliary	generate metadata	{metadata}

Training 🐧 Audio Language Model



$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{l} \log p_\gamma(c_j^i | p_1^i, ..., p_{2k}^i, c_1^i, ..., c_{j-1}^i)$$

Two types of downstream tasks

Open-ended tasks

Audio Captioning
Audio QA

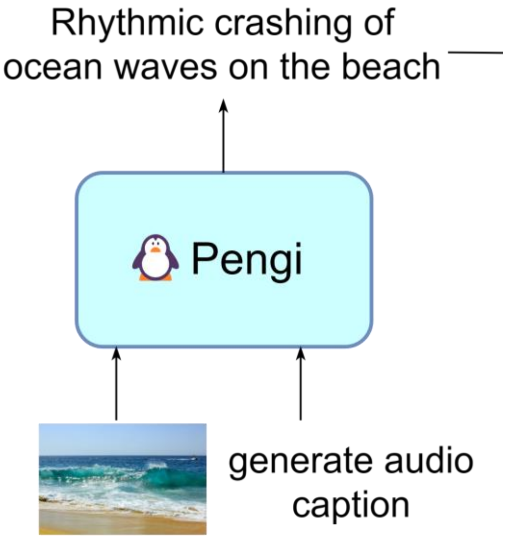
Close-ended tasks

Sound event and scene classification
Audio Retrieval
Music Analysis
Speech Emotion Recognition

Two types of downstream tasks

Open-ended tasks

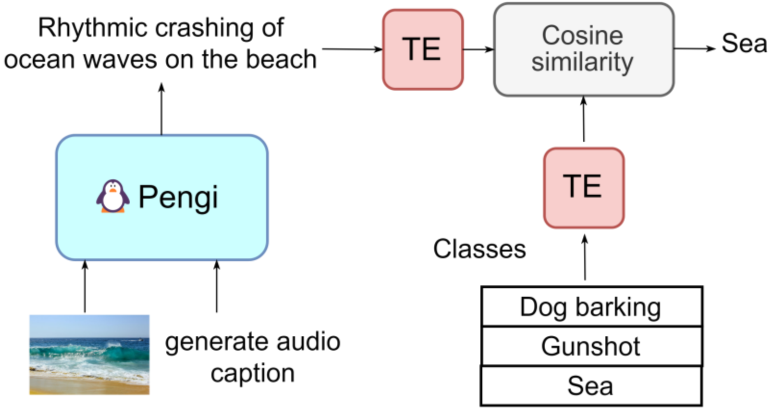
Audio Captioning
Audio QA



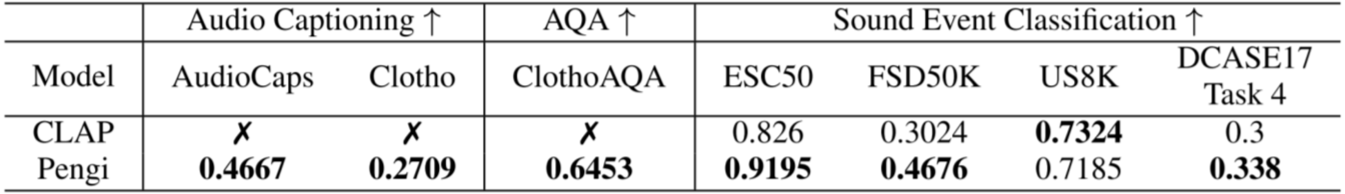
Two types of downstream tasks

Close-ended tasks

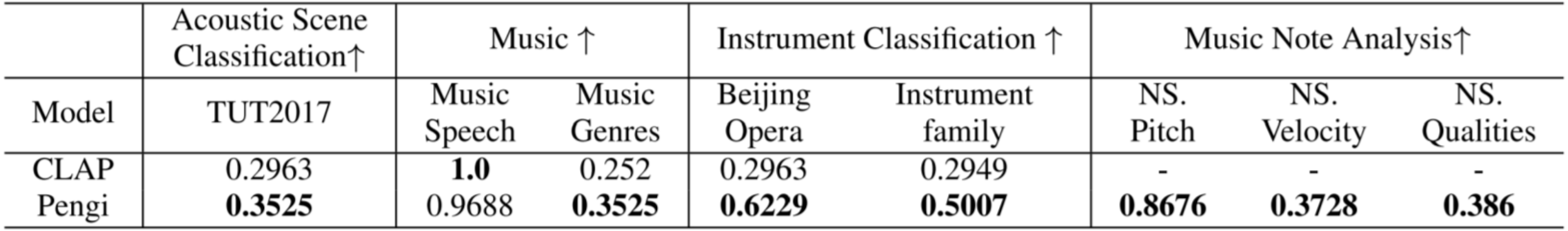
Sound event and scene classification
Text-to-Audio Retrieval
Music Analysis
Speech Emotion Recognition



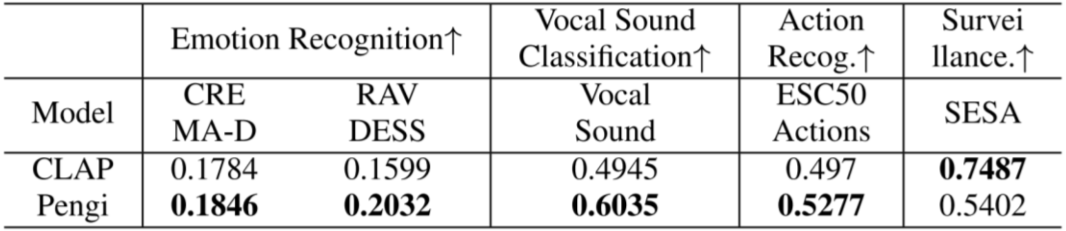
SoTA on several downstream tasks



	Audio Captioning ↑		AQA ↑	Sound Event Classification ↑			
Model	AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4
CLAP	✗	✗	✗	0.826	0.3024	**0.7324**	0.3
Pengi	**0.4667**	**0.2709**	**0.6453**	**0.9195**	**0.4676**	0.7185	**0.338**



	Acoustic Scene Classification↑	Music ↑		Instrument Classification ↑		Music Note Analysis↑		
Model	TUT2017	Music Speech	Music Genres	Beijing Opera	Instrument family	NS. Pitch	NS. Velocity	NS. Qualities
CLAP	0.2963	**1.0**	0.252	0.2963	0.2949	-	-	-
Pengi	**0.3525**	0.9688	**0.3525**	**0.6229**	**0.5007**	**0.8676**	**0.3728**	**0.386**



	Emotion Recognition↑		Vocal Sound Classification↑	Action Recog.↑	Survei llance.↑
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA
CLAP	0.1784	0.1599	0.4945	0.497	**0.7487**
Pengi	**0.1846**	**0.2032**	**0.6035**	**0.5277**	0.5402

Audio Grounded text continuation

Audio input

Text input

generate metadata

Additional text input	Text output
-	a bird is chirping
the bird is called	a robin
name the bird.	this is a bird called robin
the background is	quiet
mention forest.	a blackbird is singing in the forest

Audio input

Text input

generate metadata

Additional text input	Text output
-	a choir is singing
at the beginning a	vocalist is singing
in the end	people applaud
the background is	a sound of drum beat

NEURAL INFORMATION PROCESSING SYSTEMS

Conclusions

- Contrastive Audio-Language models are used for zero-shot close-ended tasks, such as classification and retrieval
- We propose 🐧 Pengi an Audio-Language model that can perform both open-ended and close-ended downstream tasks
- Pengi is evaluated on 21 downstream tasks and achieves SOTA performance on open-ended tasks and most close-ended tasks
- Code and pretrained models are available at https://github.com/microsoft/pengi