NATURAL LANGUAGE SUPERVISION FOR GENERAL-PURPOSE AUDIO REPRESENTATIONS

Benjamin Elizalde*, Soham Deshmukh*, Huaming Wang
Emails: {benjaminm, sdeshmukh, huawang}@microsoft.com
Microsoft

Poster in 3 sentences

1. Contrastive Language-Audio Pretraining (CLAP) connects audio and language with two encoders and a contrastive learning loss.
2. Pretrained with 4.6M audio-text pairs; the audio encoder (HTSAT) is trained on 22 audio tasks; the text encoder is a modified autoregressive decoder-only model (GPT2).
3. Our model achieves SoTA in about 26 tasks outperforming 4 different models.

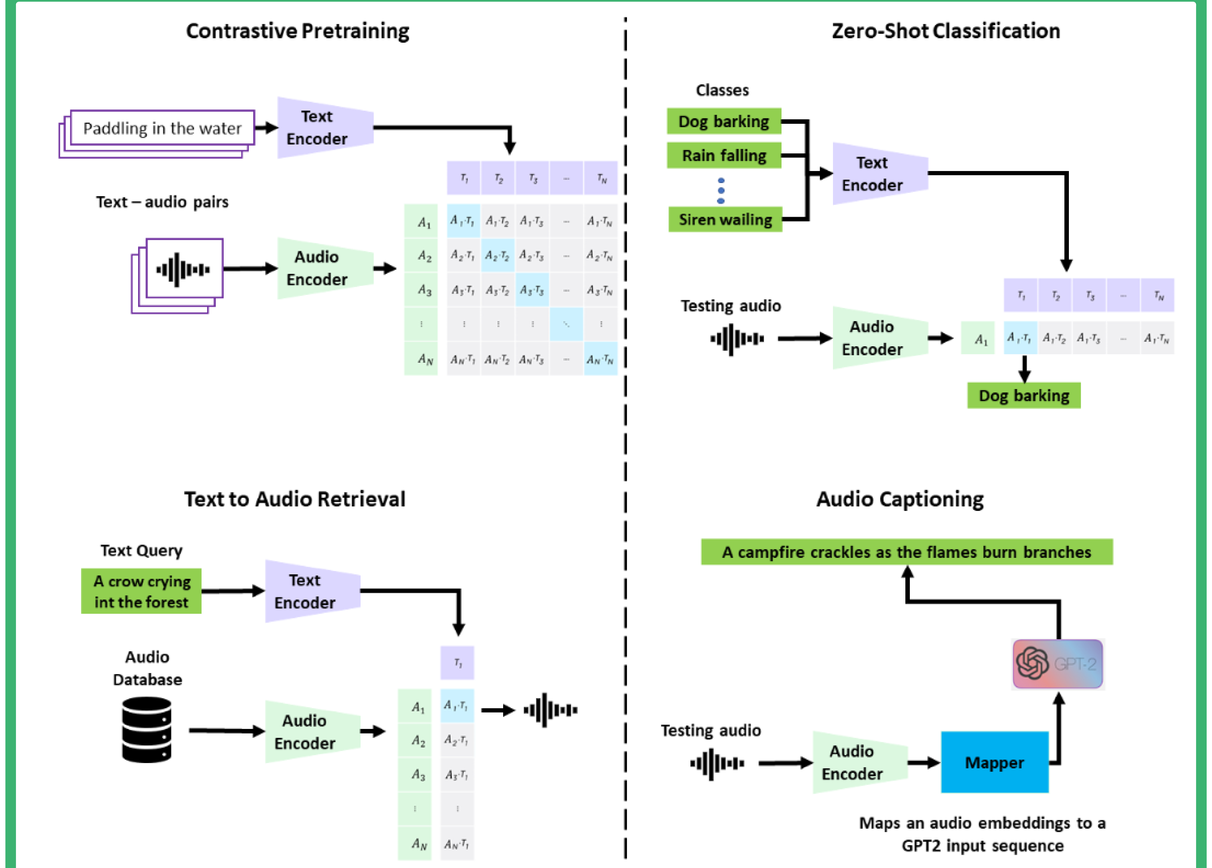
Code available on:

github.com/microsoft/CLAP

Keywords

- **Contrastive Pretraining Data:** 4.6M pairs from WavCaps, AudioSet, FSD50K, Clotho, AudioCaps, MACS, WavText5k, SoundDesc, NSynth, FMA, Mosi, Meld, Iemocap, Mosei, MSP-Podcast, CochlScene, LJspeech, EpicKitchen, Kinectics700, findsounds.com.
- **Text Encoder (GPT2)** is adapted to embed sentences by appending the special token < |endof text| > at the end of each input text.
- **Audio Encoder (HTSAT-22)** is independently trained on 22 audio tasks.
- **Contrastive Loss:** We used cosine similarity to compare text and audio; multimodal embeddings are 1024 dims: batch size 1536.
- **Audio Captioning:** At training time, only the weights of the mapper network are learned with a captioning loss.
- **Benchmark** - SoTA in the literature.

CLAP pretraining and results on 26 downstream tasks



Model	Sound Event Classification ↑					Vocal Sound Classification ↑	Surveillance Sound Classif.↑	Action Classification↑	Acoustic Scene Classification↑
	ESC50 [28]	FSD50K [12]	US8K [29]	DCASE17 Task 4 [30]	AudioSet [2]	Vocal Sound [31]	SESA [32]	ESC50 Actions [33]	TUT 2017 [30]
Benchmark	**0.948** [6]	0.302 [3]	0.806 [6]	0.3 [3]	0.058 [3]	0.495 [3]	0.25	0.045	0.296 [3]
HTSAT-22+GPT2	0.939	**0.485**	**0.823**	**0.466**	**0.102**	**0.8**	**0.65**	**0.509**	**0.538**

Model	Music Classification ↑					Instrument Classification ↑		Speech Emotion Classification↑		KWS↑	Speaker Counting↑
	GTZAN Music Speech [1]	GTZAN Genres [1]	NS Pitch [17]	NS Velocity [17]	NS Qualities [17]	Beijing Opera [1]	NS Instr. family [17]	CRE MA-D [1]	RAV DESS [34]	Speech Commands [1]	Libri Count10 [1]
Benchmark	1 [3]	0.25 [3]	0.015	0.2	0.1	**0.4746** [3]	0.09	0.178 [3]	0.159 [3]	0.106 [3]	0.178 [3]
HTSAT-22+GPT2	0.992	**0.584**	**0.444**	**0.222**	**0.489**	0.466	**0.479**	**0.3**	**0.315**	**0.164**	**0.246**

Model	Audio Captioning ↑		Audio-Text Retrieval ↑				Text-Audio Retrieval ↑			
	AudioCaps [14]	Clotho [13]	AudioCaps R@1	AudioCaps mAP@10	Clotho R@1	Clotho mAP@10	AudioCaps R@1	AudioCaps mAP@10	Clotho R@1	Clotho mAP@10
Benchmark	0.438[35]	0.215[35]	**0.517**[6]	**0.457** [4]	**0.234**[6]	0.138[4]	**0.397**[6]	0.51[4]	**0.195**[6]	0.204[4]
HTSAT-22+GPT2	**0.455**	**0.271**	0.425	0.319	0.229	**0.155**	0.356	**0.51**	0.157	**0.257**

Table 2: Performance on 26 downstream tasks using our proposed encoders and 4.6M training pairs. As the benchmark, we used the best numbers in the literature, when no number was available we used random performance. Higher is better for all tasks. The evaluation metrics are mAP for FSD50k, ESC50-Actions, AudioSet, and NS Qualities; F1-score for DCASE17; and SPIDEr for Captioning; all others use Accuracy.

Insights

1. Our proposed encoders HTSAT-22+GPT2 outperformed the best combination of encoders in the literature (Figure 1).
2. Our HTSAT-22 audio encoder is the major contributor to performance improvement.
3. Adding diversity and scaling the audio-text pairs in training presents a trade-off that increases performance in some tasks but decreases it in others.
4. Zero-Shot models should be evaluated across different domains and tasks with focus on generalization rather than on over fitting to specific tasks.
5. We only fine-tuned the audio encoder for ESC50 and achieved SoTA with 98.25% acc.

Model	Zero-Shot Score ↑ Average
CNN14+BERT	0.428
HTSAT+CLIP	0.430
HTSAT+RoBERTa	0.431
HTSAT+GPT2	0.435
HTSAT-22+RoBERTa	0.454
HTSAT-22+CLIP	0.469
HTSAT-22+GPT2	**0.480**

Figure 1:Comparison of different audio and text encoders in the literature. Zero-Shot score is the average of the metrics on 16 downstream tasks (higher is better). All models are trained on 119k training pairs. Our proposed encoders (HTSAT-22+GPT2) outperformed them all.