Tampere University

On Negative Sampling for Contrastive Audio-Text Retrieval

Huang Xie, Okko Räsänen, Tuomas Virtanen

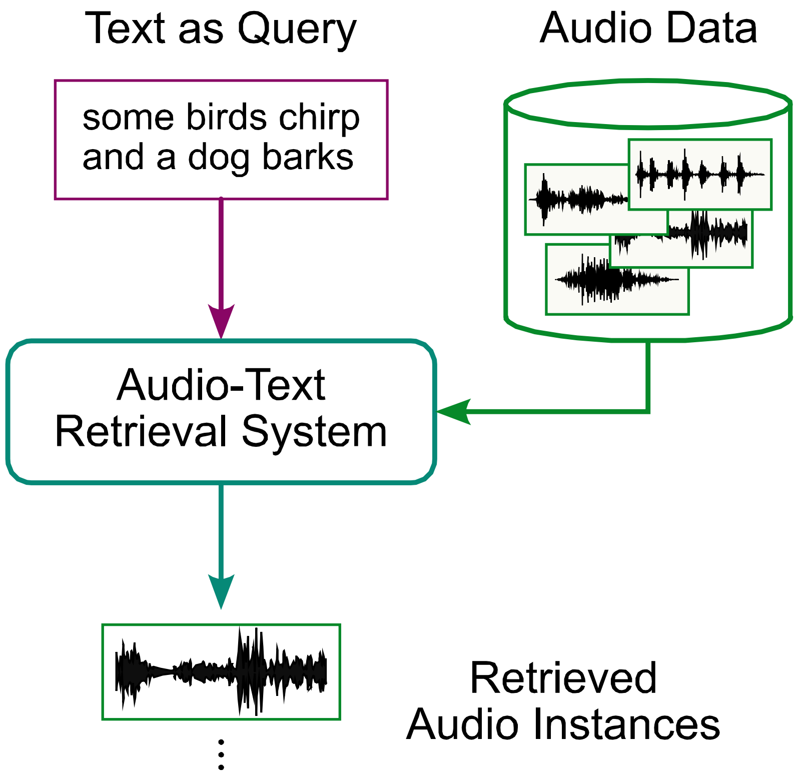
Tampere University

presented by Huang Xie

Motivation

Audio-text retrieval

- ➢ Retrieves audio or text instances relevant to a given query from the other modality.
  - ✓ E.g., audio retrieval with text queries
  - ✓ Real-world applications such as search engines



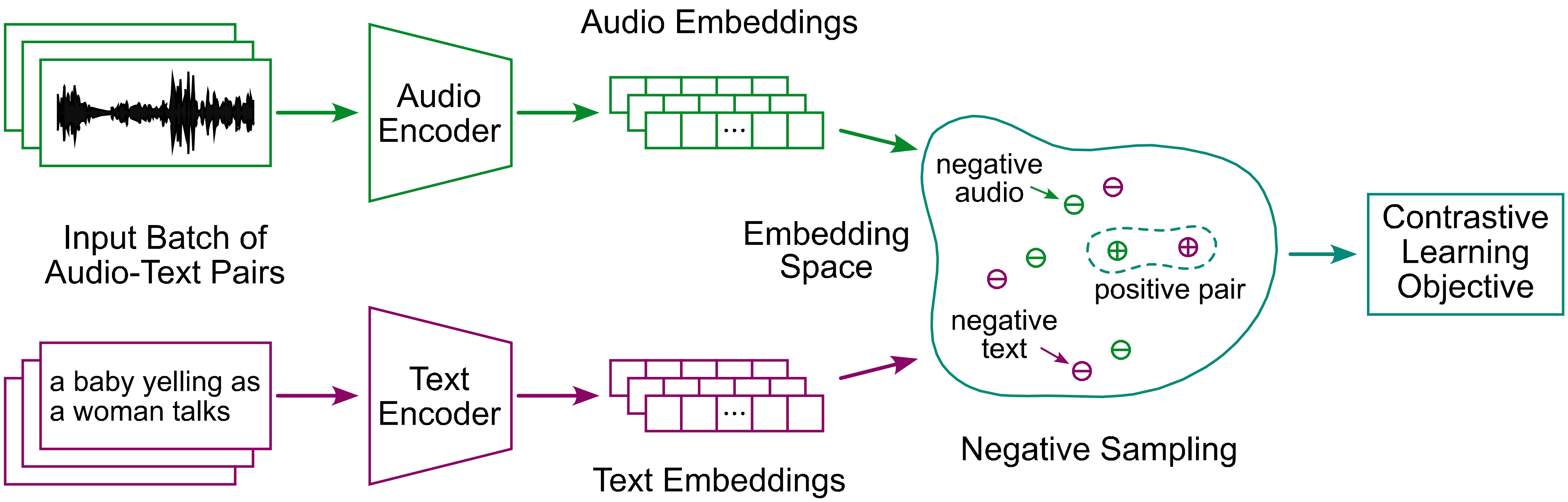
Contrastive audio-text retrieval

- ➢ Tackles audio-text retrieval with contrastive learning.

Negative sampling (NS)

- ➢ Selects informative negative samples for training.
  - ✓ Most negatives are easy to discriminate.
  - ✓ Some negatives are even counterproductive.

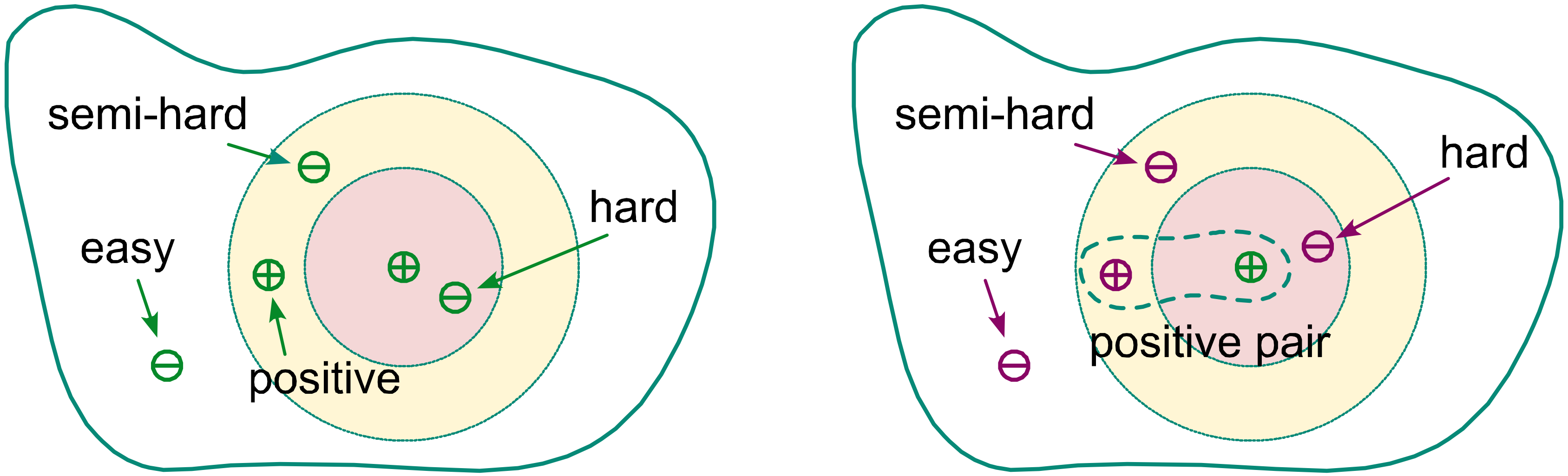
Contrastive Learning Framework



Score-based Negative Sampling

Sample hardness

- Indicates how difficult a negative sample can be distinguished from positive ones.
  - ✓ The more difficult, the more informative.
  - ✓ E.g., easy, hard, semi-hard negatives [1].



Single Modality

Two Modalities

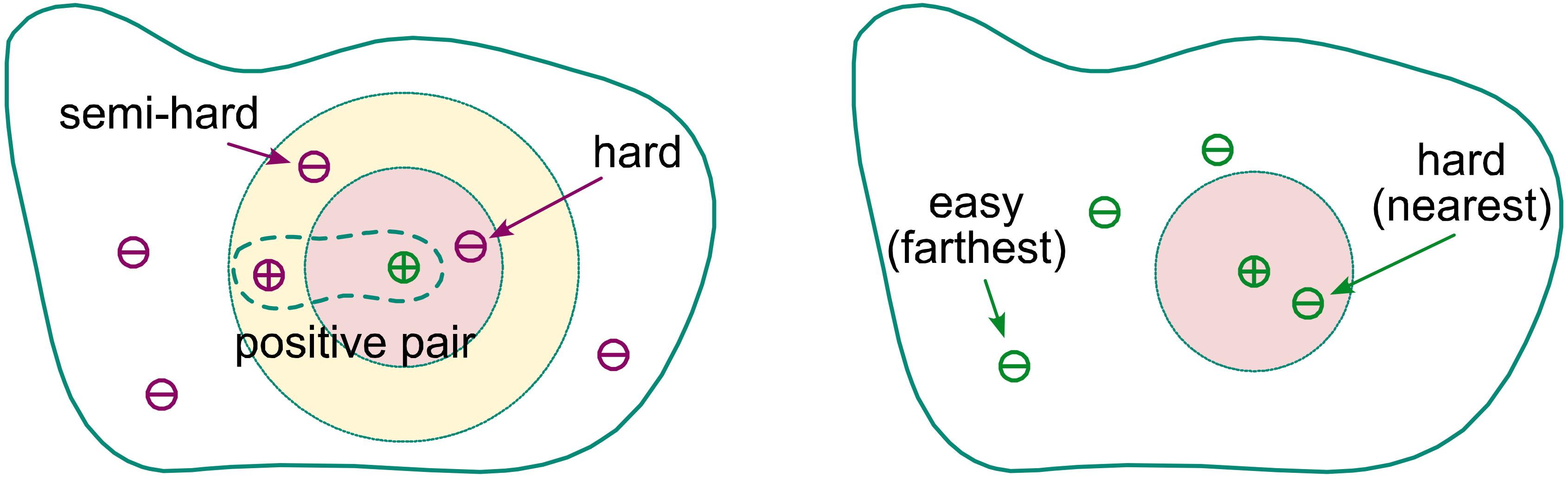
[1] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in CVPR, 2015, pp. 815-823.

Tampere University

Score-based Negative Sampling

Score-based NS

- Given a positive audio-text pair, measure sample hardness with sample similarity scores on the positive audio/text.
  - ✓ Cross-modality scores (e.g., semi-hard NS, hard NS)
  - ✓ Within-modality scores (e.g., easy NS, hard NS)



Cross Modality
(i.e., audio-text)

Within Modality
(i.e., text-based, audio-based)

Basic Negative Sampling

Random NS

- Selects negative samples at random.
- Commonly used as the default NS method.

Full-mini-batch NS

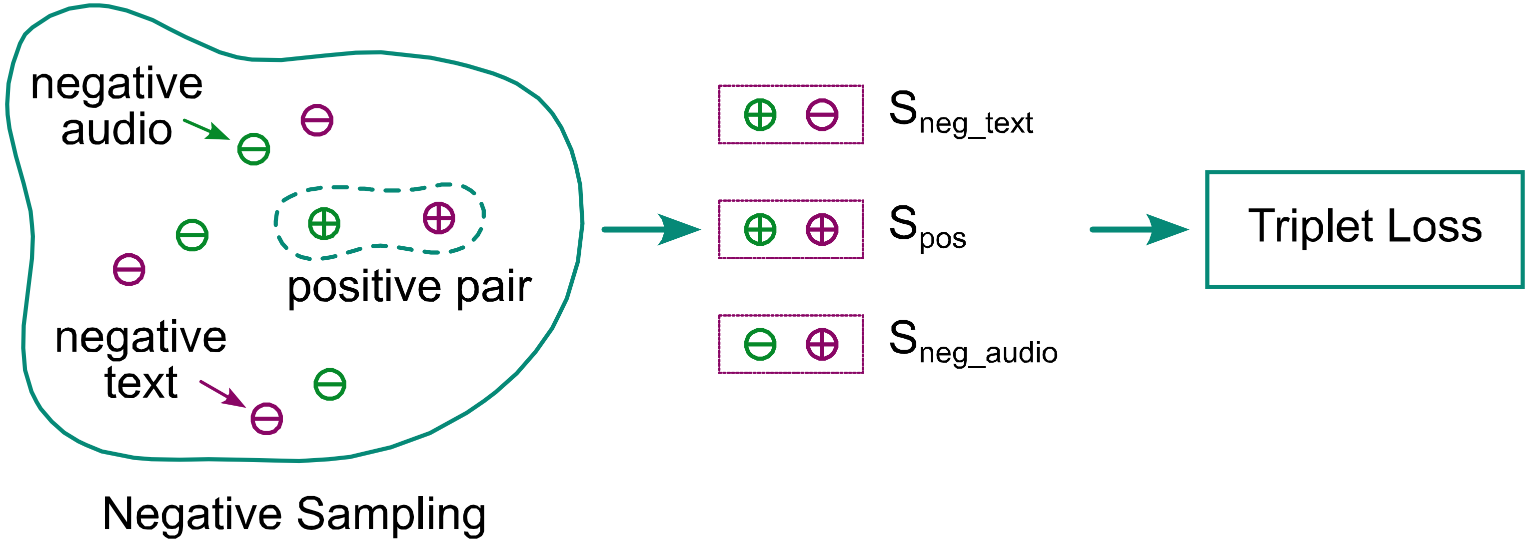
- Selects all negative samples within a mini-batch.
- Has more negative samples contributing to training.

Experiments – Contrastive Learning Objective

Triplet loss

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, S_{neg\_text} - S_{pos} + 1\right) + \max\left(0, S_{neg\_audio} - S_{pos} + 1\right)\right]$$

- N: batch size



Experiments – Audio Encoder

Convolutional recurrent neural network (CRNN) [2]

- Five convolutional blocks + one bidirectional gated recurrent unit (BiGRU)

Input

- 64-dimensional log-mel energies (40 ms frame shift)

Output

- 300-dimensional frame-level acoustic embeddings
- L2-normalized

[2] X. Xu, H. Dinkel, M. Wu, and K. Yu, “Text-to-audio grounding: Building correspondence between captions and sound events,” in ICASSP, 2021, pp. 606-610.

Experiments – Text Encoder

Word2Vec [3]

- Two-layer fully-connected neural network with the skip-gram architecture
- Pre-trained with Google News dataset (about 100 billion words)

Output

- 300-dimensional word embeddings
- L2-normalized

[3] Word2Vec, https://code.google.com/archive/p/word2vec/.

Experiments – Text & Audio Similarities

Audio-text similarity [4]

- Averaged dot products of acoustic embeddings and word embeddings

Audio-audio similarity

- Averaged dot products of acoustic embeddings

Text-text similarity

- Averaged dot products of word embeddings

[4] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, “Unsupervised Audio-Caption Aligning Learns Correspondences between Individual Sound Events and Textual Phrases,” in ICASSP, 2022, pp. 8867-8871.

Experiments – Dataset

Clotho dataset [5]

- 5,929 audio clips with a duration of 15-30 seconds
- 29,645 human written captions with a length of 8-20 words
  - ✓ Five captions for each clip
- Data splits
  - ✓ development → training
  - ✓ validation → validation
  - ✓ evaluation → evaluation

Data Split	#Clips	#Captions
development	3,839	19,195
validation	1,045	5,225
evaluation	1,045	5,225

[5] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: an Audio Captioning Dataset,” in ICASSP, 2020, pp. 736-740.

Experiments – Text-to-Audio Retrieval

Evaluation task setup

- Given a text as the query, retrieve its paired audio from 1,045 candidates.
  - ✓ One positive + 1,044 negatives

Evaluation metrics

- Mean average precision (mAP)
- Recall at rank K (R@5, R@10)

Results

- Vary dramatically
- Best with semi-hard negatives

Negative Sampling (NS)		mAP	R@5	R@10
Basic	Random NS	0.057	0.074	0.129
	Full-mini-batch NS	0.054	0.064	0.120
Score-based	**Cross-modality Semi-hard NS**	**0.121**	**0.171**	**0.274**
	Cross-modality Hard NS	0.007	0.005	0.010
	Text-based NS (hard)	0.065	0.083	0.148
	Text-based NS (easy)	0.028	0.033	0.057
	Audio-based NS (hard)	0.034	0.037	0.072
	Audio-based NS (easy)	0.011	0.005	0.010

Experiments – Audio-to-Text Retrieval

Evaluation task setup

- ➢ Given an audio clip as the query, retrieve its paired captions from 5,225 candidates.
  - ✓ Five positives + 5,220 negatives

Evaluation metrics

- ➢ Mean average precision (mAP)
- ➢ Recall at rank K (R@5, R@10)

Results

- ➢ Vary dramatically
- ➢ Best with semi-hard negatives

Negative Sampling (NS)		mAP	R@5	R@10
Basic	Random NS	0.030	0.018	0.036
	Full-mini-batch NS	0.030	0.019	0.037
Score-based	**Cross-modality Semi-hard NS**	**0.046**	**0.030**	**0.058**
	Cross-modality Hard NS	0.004	0.001	0.002
	Text-based NS (hard)	0.027	0.017	0.031
	Text-based NS (easy)	0.018	0.011	0.021
	Audio-based NS (hard)	0.030	0.018	0.035
	Audio-based NS (easy)	0.005	0.003	0.005

Conclusion

We explored score-based negative sampling by employing

- ➢ Cross-modality similarity scores
  - ✓ i.e., audio-text.
- ➢ Within-modality similarity scores
  - ✓ i.e., text-based, audio-based.

We evaluated eight negative sampling methods for contrastive audio-text retrieval.

- ➢ Six score-based methods
- ➢ Two basic methods

Thank You For Watching!

– Huang Xie