ON NEGATIVE SAMPLING FOR CONTRASTIVE AUDIO-TEXT RETRIEVAL

Huang Xie, Okko Räsänen, Tuomas Virtanen

Unit of Computing Sciences, Tampere University, Finland

ABSTRACT

This paper investigates negative sampling for contrastive learning in the context of audio-text retrieval. The strategy for negative sampling refers to selecting negatives (either audio clips or textual descriptions) from a pool of candidates for a positive audio-text pair. We explore sampling strategies via model-estimated within-modality and cross-modality relevance scores for audio and text samples. With a constant training setting on the retrieval system from [1], we study eight sampling strategies, including hard and semi-hard negative sampling. Experimental results show that retrieval performance varies dramatically among different strategies. Particularly, by selecting semi-hard negatives with cross-modality scores, the retrieval system gains improved performance in both text-to-audio and audio-to-text retrieval. Besides, we show that feature collapse occurs while sampling hard negatives with cross-modality scores.

Index Terms— Cross-modal retrieval, contrastive learning, triplet loss, negative sampling, audio-text retrieval

1. INTRODUCTION

Audio-text retrieval refers to retrieving audio or descriptive text that is relevant to a given query from the other modality. It has great potential in real-world applications, such as search engines and multimedia databases. Early works [2, 3, 4, 5] have focused on audio retrieval with separate words, for example, using words "horse trot" to search for audio containing sounds like horse trotting or galloping. Real-world audio inherently consists of sounds distributed across the temporal axis. Retrieving audio with separate words usually emphasizes on the presence of certain sounds and neglects their temporal information (e.g., relative positions on the temporal axis). A more natural way for humans to describe the desired data is by natural language descriptions. For example, a detailed description "a woman talks followed by a dog barking" provides more information than words like "human voice" and "dog barking". This paper concentrates on audio-text retrieval with unconstrained textual descriptions.

With the availability of audio-caption datasets [6, 7], several works have explored audio-text retrieval with free-form textual descriptions (i.e., audio captions). Oncescu *et al.* [8] first established benchmarks in this topic by adapting video retrieval models. Our previous work [1] investigated audio-text retrieval by learning alignment between audio and their corresponding captions. Mei *et al.* [9] evaluated the popular learning objectives (e.g., triplet losses [10, 11]) for training audio-text retrieval models. Recently, Deshmukh *et al.* [12] proposed a contrastive learning framework for audio-text retrieval, where they combined the latest advancements in audio models (e.g., CNN14 [13], HTSAT [14]) with a pretraining approach named CLAP [15]. They also introduced a new collection of audio-text pairs, which was referred to as WavText5K. Besides, the newly introduced *language-based audio retrieval* task [16] in DCASE 2022 Challenge received a total number of 31 submissions, which showed an increasing interest in this topic from the audio research community.

Most of the literature tackle audio-text retrieval with contrastive learning methods, i.e., *contrastive audio-text retrieval*. A dual-encoder architecture, consisting of an audio encoder and a text encoder, has been widely employed [16]. Audio data and textual descriptions are encoded as embeddings into a common multimodal space. By optimizing a contrastive learning objective (e.g., triplet loss), the relevant audio and text embeddings are pulled close to each other, while the irrelevant ones are pushed far away from each other in the shared space. The aforementioned works [8, 1, 9, 12] have mainly focused on the aspects such as retrieval architecture, learning objective, pretraining approach, and audio-text data collection. In contrast, the selection of negative samples for contrastive learning, i.e., negative sampling (NS), is under-explored. Previous works [11, 17] from computer vision show that the choice of negative samples has a decisive impact on the success of contrastive learning. For example, most negative samples are easy to discriminate, having minor contributions to model training [17], and some are even counterproductive, leading to a collapsed model [11].

This work aims to study and compare different strategies for negative sampling in contrastive audio-text retrieval. Particularly, we explore negative sampling via model-estimated within-modality and cross-modality relevance scores between audio and text samples in a mini-batch. With a constant training setting on the retrieval system from [1], we study eight sampling strategies, including hard and semi-hard negative sampling. Experimental results show that retrieval performance varies dramatically among these strategies. We empirically demonstrate that employing an appropriate strategy for negative sampling (e.g., selecting semi-hard negatives with cross-modality scores) can result in improved performance.

2. CONTRASTIVE AUDIO-TEXT RETRIEVAL

In this section, we formulate the task of contrastive audio-text retrieval with a model-agnostic dual-encoder framework, as illustrated in Figure 1. Let $X = \{(x_i, y_i)\}_{i=1}^{N}$ be a batch of

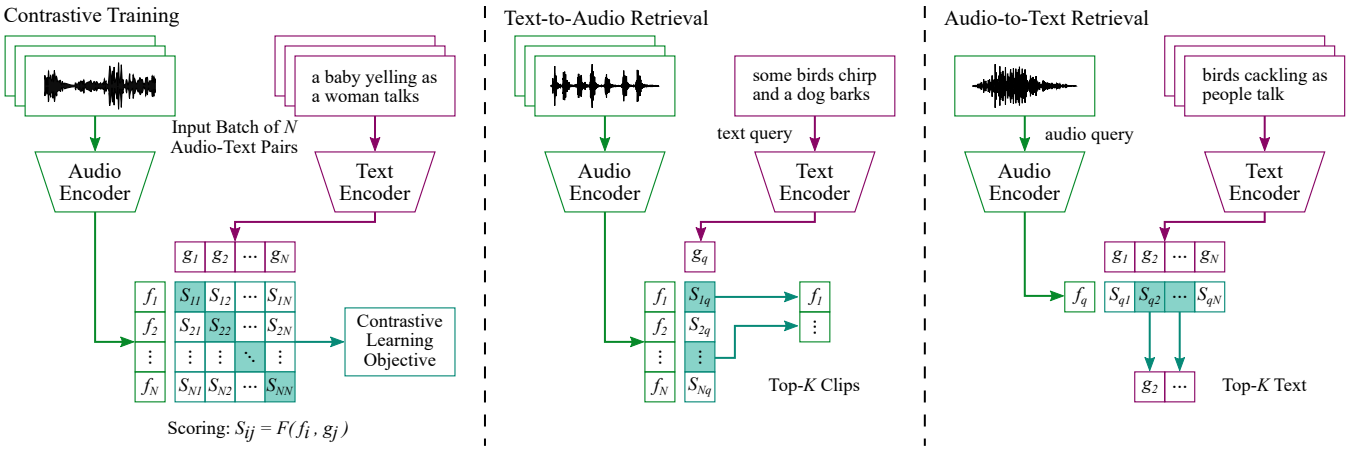

Fig. 1. Contrastive audio-text retrieval framework. From left to right: 1) Contrastive training. The dual-encoder framework is trained on a batch of N audio-text pairs by optimizing a contrastive learning objective (e.g., triplet loss). 2) Text-to-audio retrieval. For a given text query (e.g., "some birds chirp and a dog barks"), the top-K audio clips that have a high score S with the text query are retrieved. 3) Audio-to-text retrieval. For a given audio clip, the top-K textual descriptions that have a high score S with the audio query are retrieved.

N audio-text pairs, where the i-th audio clip x_i is described with the i-th textual description y_i. Due to the lack of graded audio-text relevance in existing datasets (e.g., Clotho [6], AudioCaps [7]), we consider only binary relevance between audio clips and textual descriptions. Thus, (x_i, y_i) is regarded as a positive pair and (x_i, y_j) with $i \neq j$ as a negative pair.

The audio-text pairs are projected into a common representation space via the dual-encoder framework. Let $\theta(\cdot)$ be the audio encoder, and $\phi(\cdot)$ be the text encoder. With the projected representations $f_i = \theta(x_i)$ and $g_j = \phi(y_j)$, a function $F(\cdot)$, written as

$$S_{ij} = F(f_i, g_j), \tag{1}$$

is defined to measure the semantic relevance between x_i and y_j with a score S_{ij}. The popular choices of $F(\cdot)$ include dot product [1, 12] and cosine similarity [8, 9].

The encoders $\theta(\cdot)$ and $\phi(\cdot)$ are trained on X with a contrastive learning objective $\mathcal{L}$. By optimizing $\mathcal{L}$, f_i and g_i are pulled close to each other (i.e., being of high relevance), while f_i and g_j with $i \neq j$ are pushed far away from each other (i.e., being of low relevance). The common choices of $\mathcal{L}$ include triplet loss [1, 8, 16] and symmetric cross-entropy loss [9, 12]. The triplet loss takes as input triplets of samples, consisting of one negative sample along with one positive audio-text pair, while the symmetric cross-entropy loss includes multiple negative samples for every positive pair. To alleviate the influence of sample quantity on evaluation and investigate the effect of sample quality, we utilize the instance-based triplet loss [1] (see Section 4.2). For audio-text retrieval, audio or text samples that have a high score S with the given query from the other modality are retrieved, as illustrated in Figure 1.

3. NEGATIVE SAMPLING

This section presents eight different strategies for negative sampling in contrastive audio-text retrieval. To avoid comparing positive pairs with all negative samples (i.e., heavy computational burden), we apply these strategies to select negative samples within a mini-batch of audio-text pairs, i.e., mini-batched negative sampling. These eight strategies are grouped into two categories: basic and score-based.

3.1. Basic Negative Sampling

With sampling strategies in this category, each negative sample has the same chance of being selected.

Random Negative Sampling. The simple random strategy works with a uniform distribution over negative samples in either modality, and selects one negative sample from each modality for every positive audio-text pair. It is the most basic strategy for negative sampling, and commonly used as the default setup in contrastive learning [18]. Therefore, we employ this strategy as the baseline method in this work.

Full-Mini-Batch Negative Sampling. The full-mini-batch strategy arbitrarily selects all negative samples within the same mini-batch for each positive pair. It is generally believed that contrastive learning benefits from increased sample size [19]. With this strategy, more negative samples contribute to training the dual-encoder framework.

3.2. Score-based Negative Sampling

Previous works [11, 17] show that informative negative samples promote model optimization. The term *hardness* is commonly used to represent how informative a negative sample is for a positive pair. For example, negative samples that are difficult to distinguish from positive ones are often mentioned as *hard negatives* [17], which are intuitively informative.

We define the hardness of a negative sample by its score on a positive audio-text pair. Specifically, we compute within-modality and cross-modality scores for negative audio and text samples with (1). For an audio-text pair (x_i, y_i), a text

sample y_j has a within-modality score

$$S_{ij}^{text} = F(g_i, g_j). \tag{2}$$

Similarly, an audio sample x_k has a within-modality score

$$S_{ik}^{audio} = F(f_i, f_k). \tag{3}$$

Their respective cross-modality score S_{ij} and S_{ki} are calculated directly with (1).

Within-Modality Hard Negative Sampling. This strategy treats negative samples that have the highest within-modality score, along with their paired counterpart, as hard negatives. We experiment with selecting hard negatives based on either S^{text} (i.e., text-based) or S^{audio} (i.e., audio-based). For example, in text-based hard negative sampling, the text sample y_j having the maximal S_{ij}^{text}, together with its paired audio x_j, are the hard negatives for (x_i, y_i). For comparison, we also experiment with easy negatives, i.e., negative samples that have the lowest within-modality score.

Cross-Modality Hard Negative Sampling. In this strategy, negative samples that have the highest cross-modality score on a positive pair are selected as hard negatives for experiment. For an audio-text pair (x_i, y_i), we collect its hard negatives y_j in text modality (i.e., having maximal S_{ij}) and x_k in audio modality (i.e., having maximal S_{ki}).

Cross-Modality Semi-Hard Negative Sampling. The term *semi-hard negative* was originally coined in face recognition [11]. Negative face images, which had a distance d_{neg} (e.g., squared Euclidean distance) away from the anchor similar to d_{pos} of the positive one (i.e., $d_{pos} < d_{neg} < d_{pos} + \varepsilon$ with a margin ε), were called semi-hard. We adapt this idea with cross-modality scores to select semi-hard negative samples for a positive audio-text pair. Specifically, we take negative samples that have a cross-modality score closest to that of the positive pair as semi-hard negatives. For example, an audio-text pair (x_i, y_i) has two semi-hard negatives: a text sample y_j with minimal $|S_{ij} - S_{ii}|$, and an audio sample x_k with minimal $|S_{ki} - S_{ii}|$.

4. EXPERIMENTS

This section introduces our experimental setup with the Clotho v2 dataset [6] and the retrieval system from [1].

4.1. Dataset

The Clotho v2 dataset [6] consists of 5,929 audio clips, with each clip having five human written captions (i.e., 29,645 in total), which makes it naturally suitable for audio-text retrieval. Each clip lasts for 15–30 seconds, and every caption contains 8–20 words. This dataset is divided into three splits: a development split with 3,839 clips and 19,195 captions, a validation split with 1,045 clips and 5,225 captions, and an evaluation split with 1,045 clips and 5,225 captions, respectively. All audio clips are sourced from the Freesound platform [20], and captions are crowd-sourced using a three-step framework [6]. Due to the fact of data imbalance (i.e., having more captions than audio clips), retrieving captions becomes more difficult than retrieving audio clips in this dataset.

4.2. Retrieval System

We perform audio-text retrieval with the aligning framework from [1], where a convolutional recurrent neural network (CRNN) [21] is employed as the audio encoder and a pretrained Word2Vec (Skip-gram model) [22] as the text encoder. This system is simple, trainable with triplet loss, and having few external dependencies (e.g., pretrained audio experts, external data for pretraining), which makes it convenient for experiment. Audio-text relevance scores are computed with audio frame embeddings and word embeddings [1].

Audio Encoder. The CRNN encoder [21] extracts frame-wise acoustic embeddings from audio clips of variable length. Audio clips are pre-processed using a Hanning window of 40 ms with a hop length of 20 ms. Then, 64-dimensional log mel-band energies are extracted and fed into the CRNN encoder. For each audio clip, a sequence of 300-dimensional frame embeddings are generated.

Text Encoder. Following [1], we utilize the same pretrained Word2Vec [22] as the text encoder. It includes 300-dimensional word embeddings for roughly three million case-sensitive English words. We convert captions into sequences of word embeddings in a word-by-word manner.

Triplet Ranking Loss. The retrieval system is optimized by minimizing an instance-based triplet ranking loss [1]

$$\begin{aligned}\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}[&\max(0, S_{ij} - S_{ii} + 1) \\ &+ \max(0, S_{ki} - S_{ii} + 1)],\end{aligned} \tag{4}$$

where j ($j \neq i$) indexes the sampled negative caption y_j and k ($k \neq i$) the sampled negative clip x_k for a positive audio-text pair (x_i, y_i). Audio-text relevance score S is computed by averaging trimmed dot products of frame and word embeddings [1]. For full-mini-batch NS, S_{ij} and S_{ki} are averaged over negative captions and clips of (x_i, y_i), respectively.

Training Setup. We train the retrieval system with mini-batches of 32 audio-text pairs in the development split for at most 120 epochs, and monitor the loss (4) on the validation split during training. An Adam optimizer with an initial learning rate of 0.001 is adopted to optimize the training process. The learning rate is reduced by a factor of ten once the validation loss does not improve for five epochs. Training is terminated by early stopping with a patience of ten epochs.

The trained system is used for audio-text retrieval with the evaluation split. Retrieval is performed bidirectionally between audio and text modalities: text-to-audio retrieval (i.e., retrieving audio clips that are relevant to a text query), and audio-to-text retrieval (i.e., searching for text descriptions pertaining to a given audio clip). All captions and audio clips in the evaluation split are used for retrieval.

4.3. Evaluation Metrics

Retrieval performance is measured in terms of mean average precision (mAP) and recall at k (R@k with $k \in \{5, 10\}$). The mAP is the mean of average precision (AP) over all queries,

Category	Strategy	Text-to-Audio			Audio-to-Text		
		mAP	**R@5**	**R@10**	**mAP**	**R@5**	**R@10**
Basic	Random NS	0.057	0.074	0.129	0.030	0.018	0.036
	Full-mini-batch NS	0.054	0.064	0.120	0.030	0.019	0.037
Score-based	Cross-modality Semi-hard NS	**0.121**	**0.171**	**0.274**	**0.046**	**0.030**	**0.058**
	Cross-modality Hard NS	0.007	0.005	0.010	0.004	0.001	0.002
	Text-based NS (hard)	0.065	0.083	0.148	0.027	0.017	0.031
	Text-based NS (easy)	0.028	0.033	0.057	0.018	0.011	0.021
	Audio-based NS (hard)	0.034	0.037	0.072	0.030	0.018	0.035
	Audio-based NS (easy)	0.011	0.005	0.010	0.005	0.003	0.005

Table 1. Experimental results for different negative sampling strategies with the retrieval system.

with AP being the average of precisions at positions where relevant items are in a retrieved rank list. High mAPs usually indicate that relevant items are top-ranked in the retrieval results. The R@k is defined as the proportion of relevant items among the top k retrieved results to all the relevant items contained in the evaluation data, and averaged across all queries [1, 16]. The more relevant items within top k retrieved results, the higher R@k it is. Note that, the R@k used in previous works [8, 9, 12] measures the percentage of test queries for which the correct result is among the top k retrieved results [23].

5. RESULTS AND ANALYSIS

This section presents the results for different NS strategies with a constant training setting on the retrieval system.

Overview. The mAP and R@k with $k \in \{5, 10\}$ scores obtained with different NS strategies are present in Table 1. The results show that different NS strategies have a dramatic impact on system performance. For example, in text-to-audio retrieval, "Cross-modality Semi-hard NS" (i.e., sampling semi-hard negatives with cross-modality score) achieves a mAP of 0.121 and recall scores of 0.171 (R@5) and 0.274 (R@10), which are better than those from random NS (i.e., 0.057 in mAP, 0.074 in R@5, and 0.129 in R@10). Particularly, "Cross-modality Semi-hard NS" achieves the best performance in both retrieval tasks. It obtains a mAP of 0.046 and recall scores of 0.030 (R@5) and 0.058 (R@10) in audio-to-text retrieval. Besides, all NS strategies obtain higher mAP and recall scores in text-to-audio retrieval. Since there are more captions than audio clips in the evaluation split (5,225 captions vs. 1,045 clips), audio-to-text retrieval becomes more difficult than its counterpart.

We notice that there is a gap between our results and those from previous works [9, 12, 16]. Note that, the state-of-the-art results [9, 12, 16] usually rely on heavy pretraining with external data (e.g., AudioCaps [7]), advanced retrieval architecture (e.g., CLAP [15]), and efficient learning objective (e.g., symmetric cross-entropy loss [9]). We believe that our study can contribute to these techniques to increase their performance.

Basic NS. The random NS and full-mini-batch NS obtain similar results in both retrieval tasks, with mAP / R@5 / R@10 around 0.054 / 0.064 / 0.120 in text-to-audio retrieval and 0.030 / 0.018 / 0.036 in audio-to-text retrieval. Theoretically, more negatives are involved in training the retrieval system with full-mini-batch NS, which would have an impact on the performance. A possible explanation for this could be due to the small batch size (i.e., 32) used for experiment.

Score-based NS. In contrast to "Cross-modality Semi-hard NS", "Cross-modality Hard NS" (i.e., selecting hard negatives with cross-modality score) performs worse in both tasks. For example, in text-to-audio retrieval, it obtains a mAP of 0.007 and recall scores of 0.005 (R@5) and 0.010 (R@10), even worse than those from random NS. The outputs from the audio encoder show that feature collapse occurs with "Cross-modality Hard NS" (i.e., outputting zero vectors as acoustic embeddings), which would be caused by the high variance of gradients with too hard negatives. The remaining four strategies select either hard or easy negatives with within-modality score (i.e., text-based and audio-based). The results indicate that selection of within-modality hard negatives is more efficient than the use of easy negatives.

6. CONCLUSION

In this work, we evaluated eight different negative sampling strategies for contrastive audio-text retrieval with a constant training setting on the retrieval system from [1]. The experimental results show that retrieval performance varies dramatically among different strategies. Particularly, by selecting semi-hard negatives with model-estimated cross-modality scores, the system achieves improved performance in both text-to-audio and audio-to-text retrieval. We also notice that feature collapse occurs while sampling hard negatives with cross-modality scores. Additionally, we demonstrated that hard negatives selected with within-modality scores (i.e., text-based and audio-based) are more informative than those easy ones. As future work, we consider exploring non-binary audio-text relevance and pretraining for audio-text retrieval.

7. ACKNOWLEDGEMENT

The research leading to these results has received funding from Emil Aaltonen foundation funded project "Using language to interpret unstructured data" and Academy of Finland grant no. 314602. We would like to thank Khazar Khorrami for her useful discussions.

8. REFERENCES

[1] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised Audio-Caption Aligning Learns Correspondences between Individual Sound Events and Textual Phrases," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2022, pp. 8867–8871.

[2] M. Slaney, "Semantic-Audio Retrieval," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2002, pp. IV–4108–IV–4111.

[3] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-Scale Content-Based Audio Retrieval from Text Queries," in *Proc. Int. Conf. Multimed. Inf. Retr. (MIR)*, 2008, pp. 105–112.

[4] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," in *Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE)*, 2018, pp. 59–63.

[5] B. Elizalde, S. Zarar, and B. Raj, "Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2019, pp. 4095–4099.

[6] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an Audio Captioning Dataset," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2020, pp. 736–740.

[7] C.D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating Captions for Audios in The Wild," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (HLT-NAACL)*, 2019, pp. 119–132.

[8] A.M. Oncescu, A.S. Koepke, J.F. Henriques, Z. Akata, and S. Albanie, "Audio Retrieval with Natural Language Queries," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2021, pp. 2411–2415.

[9] X. Mei, X. Liu, J. Sun, M.D. Plumbley, and W. Wang, "On Metric Learning for Audio-Text Cross-Modal Retrieval," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2022, pp. 4142–4146.

[10] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification Using a "Siamese" Time Delay Neural Network," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 1993, pp. 737–744.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. Conf. Comput. Vision and Pattern Recognit. (CVPR)*, 2015, pp. 815–823.

[12] S. Deshmukh, B. Elizalde, and H. Wang, "Audio Retrieval with WavText5K and CLAP Training," arXiv preprint arXiv:2209.14275, 2022.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M.D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, pp. 2880–2894, 2020.

[14] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *Proc. Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2022, pp. 646–650.

[15] B. Elizalde, S. Deshmukh, M.A. Ismail, and H. Wang, "CLAP: Learning Audio Concepts from Natural Language Supervision," arXiv preprint arXiv:2206.04769, 2022.

[16] H. Xie, S. Lipping, and T. Virtanen, "Language-based Audio Retrieval Task in DCASE 2022 Challenge," arXiv preprint arXiv:2206.06108, 2022.

[17] J. Robinson, C. Chuang, S. Sra, and S. Jegelka, "Contrastive Learning with Hard Negative Samples," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021, pp. 1–28.

[18] L. Xu, J. Lian, W.X. Zhao, M. Gong, L. Shou, D. Jiang, X. Xie, and J.R. Wen, "Negative Sampling for Contrastive Representation Learning: A Review," arXiv preprint arXiv:2206.00212, 2022.

[19] P. Awasthi, N. Dikkala, and P. Kamath, "Do More Negative Samples Necessarily Hurt In Contrastive Learning?," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 1101–1116.

[20] F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *Proc. Int. Conf. Multimed.*, 2013, pp. 411–412.

[21] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A CRNN-GRU Based Reinforcement Learning Approach to Audio Captioning," in *Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE)*, 2019, pp. 225–229.

[22] "Word2Vec," https://code.google.com/archive/p/word2vec, Accessed: 2021-09-27.

[23] N.C. Mithun, J. Li, F. Metze, and A.K. Roy-Chowdhury, "Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval," in *Proc. Int. Conf. Multimed. Retr. (ICMR)*, 2018, pp. 19–27.