On Metric Learning for Audio-Text Cross-Modal Retrieval

Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK
{x.mei, xubo.liu, jianyuan.sun, m.plumbley, w.wang}@surrey.ac.uk

Abstract

Audio-text retrieval aims at retrieving a target audio clip or caption from a pool of candidates given a query in another modality. Solving such cross-modal retrieval task is challenging because it not only requires learning robust feature representations for both modalities, but also requires capturing the fine-grained alignment between these two modalities. Existing cross-modal retrieval models are mostly optimized by metric learning objectives as both of them attempt to map data to an embedding space, where similar data are close together and dissimilar data are far apart. Unlike other cross-modal retrieval tasks such as image-text and video-text retrievals, audio-text retrieval is still an unexplored task. In this work, we aim to study the impact of different metric learning objectives on the audio-text retrieval task. We present an extensive evaluation of popular metric learning objectives on the AudioCaps and Clotho datasets. We demonstrate that NT-Xent loss adapted from self-supervised learning shows stable performance across different datasets and training settings, and outperforms the popular triplet-based losses. Our code is available at https://github.com/XinhaoMei/audio-text_retrieval.

Index Terms: metric learning, audio retrieval, text-based retrieval, cross-modal task

1. Introduction

Given an audio clip or a caption as a query, audio-text retrieval aims at retrieving a paired item from a pool of candidates in another modality. This cross-modal retrieval task is challenging as it requires not only learning robust feature representations for both acoustic and textual modalities but also capturing fine-grained interaction between the learned acoustic and textual features and aligning them in a shared embedding space. Audio-text retrieval can be potentially applied to applications such as film, audio book production and web search.

Cross-modal retrieval tasks (e.g. image-text retrieval and video-text retrieval) have received extensive attention in recent years and have made great progress [1, 2, 3, 4, 5, 6]. However, little attention has been paid to audio-text retrieval in the literature. One reason might be the lack of appropriate datasets, thus early works just focused on tag-based audio retrieval, where the queries were words not sentences. Chechik et al. [7] proposed a tag-based audio retrieval system using traditional machine learning techniques (e.g. support vector machines and Gaussian mixture models). Ikawa et al. [8] investigated searching sounds using onomatopoeic words. Elizalde et al. [9] employed a siamese network to align audio and textual features to a joint embedding space. Although these tag-based audio retrieval works show reasonable performance, they are constrained in the query format. Retrieving audio clips using free-form language (sentences) is more natural for humans.

With the fast development of audio captioning in recent three years [10, 11, 12, 13], publicly available audio captioning datasets are released [14, 15], which are naturally suited for the free-form language-based audio-text retrieval task. Koepke et al. [16] first established the benchmarks for free-form language-based audio retrieval, where they adapted models from video retrieval and made use of pre-trained models to solve the data scarcity problem. Since both the audio and text (captions) are sequence data, free-form language-based audio-text retrieval is more challenging than tag-based audio retrieval and is the focus of this paper. We use the term audio-text and audio-caption interchangeably in this paper.

Similar to other cross-modal retrieval models [1], the audio-text retrieval models can be built with two sub-networks, namely, an audio encoder and a text encoder. The objectives of these two encoders are to map the audios and texts into a joint embedding space, where the semantically similar embeddings are close to each other and the dissimilar items are far away. We refer to the embeddings as Acoustic Semantic Embeddings (ASE) as they are learned via jointly modeling the audio and language modalities. The training objective of cross-modal retrieval models is consistent with that of metric learning [17]. To this end, metric learning has been a popular choice for the optimization of the cross-modal retrieval models. Numerous metric learning objectives have been introduced for various tasks such as face identification [18], speaker recognition [19], and retrieval [1, 2], however, there is no clear argument which one is the most suited since they may work well on specific tasks or data but may not generalize well to other tasks [20]. In this work, we aim to study and compare the impact of different metric learning objectives for the free-form language-based audio-text retrieval task in a constant training setting.

We focus on triplet loss and its variants, as they have demonstrated promising performance and are popularly employed [18]. In a triplet setting for audio-text retrieval, an audio clip and its corresponding caption are regarded as an anchor and a positive example, respectively, while other unpaired captions are regarded as negatives. The hinge-based triplet ranking loss sums over all negative samples within a mini-batch (thus we refer to it as triplet-sum). Faghri et al. [1] argued that hard negatives should be emphasised as other easy negatives may dominate the loss and create local minimal, thus they proposed a triplet ranking loss with hard negative mining (we refer to it as triplet-max) which focuses only on the hardest negative within a mini-batch. Wei et al. [21] further proposed a universal weighting framework for cross-modal retrieval, where the pairs are weighted based on their similarity scores (we refer to it as triplet-weighted). In addition to the triplet-based losses, we further adapt a contrastive loss used in self-supervised learning here for supervised cross-modal retrieval, that is, normalized temperature-scaled cross entropy loss (NT-Xent) [22]. The NT-Xent loss is based on softmax and aims to identify the positive pairs within a mini-batch.

In summary, we first establish a baseline using pre-trained

models to learn the acoustic semantic embeddings, then we present an extensive evaluation of the popular metric learning objectives described above on our baseline in a constant training setting. Against popular belief [1, 20], we demonstrate that triplet losses with hard negative mining are sensitive to the training settings and may be hard to converge, the NT-Xent loss shows stable performance with respect to different datasets and training settings and outperforms the triplet-based losses.

2. Audio-Text Retrieval with Metric Learning

In this section, we first formulate the audio-text retrieval problem and introduce the baseline model, then the metric learning objectives (loss functions) we evaluated are introduced.

2.1. Problem formulation

Let $D = \{(a_i, t_i)\}_{i=1}^N$ be an audio captioning dataset of N examples, where a_i is an audio clip and t_i is the paired caption. Therefore, (a_i, t_i) is regarded as a positive pair while $(a_i, t_{j,j\neq i})$ is a negative pair. One audio clip could have multiple paired captions, we just consider one here for simplicity. The audio-text retrieval models usually consist of an audio encoder f and a text encoder g, which project the audio clip and text into a shared embedding space, respectively. For an audio-caption pair (a_i, t_j), the similarity of the audio and caption can be measured by cosine similarity of their embeddings:

$$s_{ij} = \frac{f(a_i) \cdot g(t_j)}{||f(a_i)||_2 ||g(t_j)||_2}. \quad (1)$$

The two encoders are trained to make the similarity score of positive pairs s_{ii} higher than that of negative pairs s_{ij}.

2.2. Model

The data available in the audio captioning datasets is limited, and this data scarcity problem usually limits the performance of the model in learning robust feature representations. Transfer learning has been adopted as a standard recipe to alleviate the data scarcity problem and has shown promising performance in audio captioning task [23]. Therefore, pre-trained models are employed here.

Audio Encoder Pre-trained audio neural networks (PANNs) [24] are pre-trained on AudioSet with an audio tagging task, which are shown to provide robust audio representations and perform well on a wide range of audio-related tasks. The ResNet-38 in PANNs is employed as the audio encoder, where the last two linear layers are discarded. An average and max pooling layer is applied to aggregate along the frequency dimension on the feature map output by the last convolutional block. A multilayer perceptron (MLP) block is used to project the audio features into a shared embedding space, which consists of two linear layers with a ReLU [25] activation layer between them.

Text Encoder Numerous large-scale per-trained language models have been published in recent year, which show powerful capability to model the language and produce contextual-aware embeddings. BERT [26], which stands for Bidirectional Encoder Representations from Transformers, obtains state-of-the-art results on a wide variety of Natural Language Processing (NLP) tasks. The pre-trained BERT is employed as the text encoder here. A "$<$CLS$>$" token is appended at the start of each sentence and used as the final sentence representation. A MLP block is also applied to project the sentence representation into the shared embedding space.

2.3. Loss functions

During training, we sample a mini-batch of audio-caption pairs $\{a_i, t_i\}_{i=1}^B$ where B is the batch size. Triplet-based loss functions are based on the concept of triplet, which is made up by an anchor, a positive (paired candidate in another modality) and a negative (unpaired candidate in another modality). The anchor with its positive is a positive pair and the anchor with its negative is a negative pair as we defined above.

Triplet-sum For each query, the triplet-sum loss aims to maximize the similarity score of its positive pair while minimizing the similarity scores to all other negatives within a mini-batch, thus it can be formulated as

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \sum_{j \neq i} [m + s_{ij} - s_{ii}]_+ + [m + s_{ji} - s_{ii}]_+, \quad (2)$$

where $[x]_+ = \max(0, x)$ and m is a distance margin. Since audio-text retrieval is a bidirectional retrieval task (audio-to-text and text-to-audio), the loss has two terms, the first term sums over all negative captions given a query audio clip while the second term sums over all negative audio clips given a query caption. If the similarity of the positive pair is larger than that of any negatives in the mini-batch by the margin m, the loss will be zero.

Triplet-max Triplet-sum loss sums over all negatives for each query within a mini-batch, Faghri et al. [1] argued that the easy negatives may dominate the loss and make it stuck into local minimal, thus, hard negatives should be emphasized. They proposed the triplet-max loss that focuses on the hardest negatives during training, which can be formulated as:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \max_{j \neq i} [m + s_{ij} - s_{ii}]_+ + \max_{j \neq i} [m + s_{ji} - s_{ii}]_+. \quad (3)$$

For each query, it aims to maximize the similarity score of its positive pair while just minimizing the similarity score to its hardest negative within a mini-batch, that is, the negative closest to the query in the embedding space. In this way, the easy negatives won't violate the loss and the hardest negative takes all the gradients.

Triplet-weighted Both the positive and negative pairs are treated equally in the triplet-sum and triplet-max losses. Wei et al. [21] further introduced a universal function $G(\cdot)$ to weight the pairs based on their similarity scores. Specifically, the weighting function is defined as a polynomial function. For a positive pair (a_i, t_i), the positive weight function G_{pos} is defined as:

$$G_{\text{pos}} = a_p s_{ii}^p + a_{p-1} s_{ii}^{p-1} + \cdots + a_1 s_{ii}^1 + a_0, \quad (4)$$

where p is the order of the polynomial function and $\{a_i\}_{i=0}^p$ are the hyper-parameters. The negative weight function G_{neg} can be formulated as:

$$G_{\text{neg}} = b_q s_{ij}^q + b_{q-1} s_{ij}^{q-1} + \cdots + b_1 s_{ij}^1 + b_0, \quad (5)$$

where $\{b_i\}_{i=0}^q$ are the hyper-parameters and q is the order. If the similarity score of the positive pair (a_i, t_i) increases, the positive weight value will decrease. In contrast, for a negative pair (a_i, t_j), the negative weight value increases if the similarity score of the negative pair increases.

Let $N_{a_i} = \{s_{ij,j\neq i}\}$ and $N_{t_j} = \{s_{ji,i\neq j}\}$ be the similarity scores for all negative pairs of an audio sample a_i and a text

Table 1: *Results of the experiments. * indicates the learning rate is 5×10^{-5} otherwise 1×10^{-4}.*

Dataset	Fine-tune	Objective	Text-to-Audio			Audio-to-Text		
			R@1	R@5	R@10	R@1	R@5	R@10
AudioCaps	No	Triplet-sum	$17.2_{\pm0.5}$	$47.6_{\pm0.2}$	$64.3_{\pm0.1}$	$19.3_{\pm1.5}$	$51.4_{\pm0.8}$	$67.2_{\pm0.7}$
		Triplet-max	$19.9_{\pm0.2}$	$51.6_{\pm0.3}$	$67.4_{\pm0.3}$	$21.2_{\pm0.9}$	$54.9_{\pm1.4}$	$70.2_{\pm1.0}$
		Triplet-weighted	$19.9_{\pm0.2}$	$52.1_{\pm0.7}$	$67.6_{\pm0.6}$	$21.9_{\pm0.8}$	$56.3_{\pm0.3}$	$71.5_{\pm0.4}$
		NT-Xent	$19.2_{\pm0.4}$	$51.1_{\pm0.2}$	$66.6_{\pm0.2}$	$21.3_{\pm0.5}$	$53.6_{\pm0.7}$	$69.8_{\pm0.5}$
	Yes	Triplet-sum	$32.2_{\pm0.3}$	$68.2_{\pm0.6}$	$81.6_{\pm0.5}$	$36.1_{\pm1.2}$	$69.2_{\pm1.3}$	$81.4_{\pm1.7}$
		Triplet-max	$32.7_{\pm0.3}$	$68.3_{\pm0.8}$	$81.6_{\pm0.5}$	$38.7_{\pm1.0}$	$70.6_{\pm0.7}$	$82.2_{\pm0.8}$
		Triplet-weighted	$32.6_{\pm0.6}$	$67.7_{\pm0.7}$	$81.0_{\pm0.8}$	$39.6_{\pm1.1}$	$72.0_{\pm2.2}$	$82.2_{\pm1.4}$
		NT-Xent	$33.9_{\pm0.4}$	$69.7_{\pm0.2}$	$82.6_{\pm0.3}$	$39.4_{\pm1.0}$	$72.0_{\pm1.0}$	$83.9_{\pm0.6}$
Clotho	No	Triplet-sum	$7.0_{\pm0.2}$	$23.0_{\pm0.3}$	$34.8_{\pm0.5}$	$8.3_{\pm0.3}$	$25.4_{\pm0.8}$	$36.7_{\pm0.5}$
		Triplet-max	$7.9_{\pm0.2}$	$23.6_{\pm0.4}$	$34.2_{\pm0.5}$	$8.8_{\pm1.0}$	$25.7_{\pm1.0}$	$36.3_{\pm0.8}$
		Triplet-weighted*	$7.2_{\pm0.2}$	$22.1_{\pm0.3}$	$33.0_{\pm0.3}$	$7.5_{\pm0.4}$	$23.2_{\pm0.1}$	$33.3_{\pm0.3}$
		NT-Xent	$8.0_{\pm0.2}$	$25.3_{\pm0.1}$	$36.9_{\pm0.3}$	$9.2_{\pm0.7}$	$27.9_{\pm0.5}$	$39.0_{\pm0.7}$
	Yes	Triplet-sum	$14.2_{\pm0.5}$	$36.6_{\pm0.5}$	$49.3_{\pm0.7}$	$16.1_{\pm0.7}$	$37.5_{\pm1.2}$	$50.7_{\pm1.0}$
		Triplet-max*	$14.2_{\pm0.9}$	$36.5_{\pm1.2}$	$49.0_{\pm0.4}$	$15.8_{\pm0.4}$	$36.4_{\pm1.6}$	$49.6_{\pm2.0}$
		Triplet-weighted*	$14.2_{\pm0.4}$	$36.6_{\pm0.5}$	$49.7_{\pm0.3}$	$16.9_{\pm0.4}$	$38.1_{\pm0.2}$	$51.4_{\pm0.2}$
		NT-Xent	$14.4_{\pm0.4}$	$36.6_{\pm0.2}$	$49.9_{\pm0.2}$	$16.2_{\pm0.7}$	$37.5_{\pm0.9}$	$50.2_{\pm0.7}$

sample t_j respectively. The loss can be formulated as:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} [\sum_{p=0}^{P} a_p s_{ii}^p + \sum_{q=0}^{Q} b_q \max\{N_{a_i}^q\}]_+$$
$$+ \frac{1}{B} \sum_{i=1}^{B} [\sum_{p=0}^{P} a_p s_{ii}^p + \sum_{q=0}^{Q} b_q \max\{N_{t_i}^q\}]_+, \quad (6)$$

where P and Q are the highest power of positive and negative pairs, respectively. The $\max$ term in the equation will select the hardest negative pair, thus the loss is referred to as the maximum polynomial loss. They also proposed another average polynomial loss that will first select informative negative pairs based on a mining policy and average the similarity scores of the selected pairs. We just employ the maximum polynomial loss here as it performs better than the average one in [21].

NT-Xent NT-Xent is a contrastive loss based on softmax, proposed by Chen et al. [22] to learn visual representations via self-supervised learning. We adapt it here for the supervised cross-modal retrieval task, as follows:

$$\mathcal{L} = -\frac{1}{B} \left(\sum_{i=1}^{B} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)} + \right.$$
$$\left. \sum_{i=1}^{B} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)} \right), \quad (7)$$

where τ is a temperature hyper-parameter. It aims to maximize the similarity of the positive pair with respect to all negative pairs within a mini-batch, and the final loss is also computed in a bidirectional manner.

3. Experiments

3.1. Datasets

We focus on free-form language-based audio-text retrieval, thus audio captioning datasets are naturally employed, in which each audio clip is annotated by humans using natural sentences. Extensive experiments are carried out on AudioCaps [14] and Clotho [15] datasets, respectively.

AudioCaps AudioCaps is the largest audio captioning dataset with around 50k audio clips. All the audio clips are 10-second long and are sourced from AudioSet [27], the largest dataset for audio tagging. In our downloaded version, the training set contains 49 274 audio clips and each of them has one corresponding human-annotated caption, the validation and test set contain 494 and 957 audio clips, respectively, and each of them has five human-annotated captions.

Clotho Clotho is an audio captioning dataset whose audio clips are collected from the Freesound[1] archive. The length of the audio clips ranges uniformly from 15 to 30 seconds. Clotho v2 is used here. There are 3839 audio clips in the training set and 1045 audio clips in the validation and test sets respectively. All the audio clips have five diverse human-annotated captions of eight to 20 words in length.

3.2. Experimental setups

Log mel-spectrograms are used as the audio features, which are extracted using a 1024-points Hanning window with 320-points hop size and 64 mel bins. All the models are trained for 50 epochs using Adam [28]. The learning rate is set to 1×10^{-4} or 5×10^{-5} and is decayed to 1/10 of itself every 20 epochs. The batch size is set to 32 for the AudioCaps dataset and 24 for the Clotho dataset. We perform experiments by freezing and fine-tuning the pre-trained models. The best model is selected based on the sum of recalls on the validation set.

For all triplet-based losses, the distance margin m is set to 0.2, and the temperature hyper-parameter τ in the NT-Xent loss is set to 0.07. For triplet-weighted loss, we follow the hyper-parameter settings in [21], that is, $Q = 2, \{a_0 = 0.5, a_1 = -0.7, a_2 = 0.2\}$, and $P = 2, \{b_0 = 0.03, b_1 = -0.4, a_2 = 0.9\}$. The dimension of the shared embedding space is 1024. The learned acoustic semantic embeddings are normalized. All experiments are carried out on a single RTX3090 GPU.

3.3. Evaluation protocol

Recall at rank k (R@k) is used as the evaluation metric, which is the popular cross-modal retrieval evaluation protocol. R@k mea-

---

[1]https://freesound.org/

Table 2: *Experimental results with different batch sizes on AudioCaps dataset. The pre-trained encoders are not fine-tuned.*

Batch size	Objective	Text-to-Audio			Audio-to-Text		
		R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
32	Triplet-sum	$17.2_{\pm 0.5}$	$47.6_{\pm 0.2}$	$64.3_{\pm 0.1}$	$19.3_{\pm 1.5}$	$51.4_{\pm 0.8}$	$67.2_{\pm 0.7}$
	Triplet-max	$19.9_{\pm 0.2}$	$51.6_{\pm 0.3}$	$67.4_{\pm 0.3}$	$21.2_{\pm 0.9}$	$54.9_{\pm 1.4}$	$70.2_{\pm 1.0}$
	Triplet-weighted	$19.9_{\pm 0.2}$	$52.1_{\pm 0.7}$	$67.6_{\pm 0.6}$	$21.9_{\pm 0.8}$	$56.3_{\pm 0.3}$	$71.5_{\pm 0.4}$
	NT-Xent	$19.2_{\pm 0.4}$	$51.1_{\pm 0.2}$	$66.6_{\pm 0.2}$	$21.3_{\pm 0.5}$	$53.6_{\pm 0.7}$	$69.8_{\pm 0.5}$
128	Triplet-sum	$17.0_{\pm 0.7}$	$47.2_{\pm 0.2}$	$62.7_{\pm 0.2}$	$20.0_{\pm 1.2}$	$49.9_{\pm 1.5}$	$66.7_{\pm 1.0}$
	Triplet-max	$11.8_{\pm 0.3}$	$38.1_{\pm 0.9}$	$53.7_{\pm 0.5}$	$15.0_{\pm 0.5}$	$43.4_{\pm 0.6}$	$59.6_{\pm 0.7}$
	Triplet-weighted	$10.1_{\pm 0.6}$	$31.7_{\pm 0.9}$	$46.5_{\pm 0.9}$	$10.7_{\pm 0.9}$	$33.4_{\pm 1.7}$	$48.7_{\pm 2.6}$
	NT-Xent	$19.5_{\pm 0.1}$	$50.4_{\pm 0.3}$	$65.6_{\pm 0.8}$	$22.2_{\pm 0.7}$	$53.7_{\pm 2.0}$	$69.5_{\pm 0.4}$

sures the percentage of targets retrieved within the top k ranked results, thus the higher the score, the better the performance. We report R@1, R@5, and R@10. All the experiments are repeated three times with different training seeds and we report the mean and standard deviation of the metrics.

4. Results

4.1. Model performance

The experimental results are shown in Table 1. The baseline model is simple but effective. It can be observed that fine-tuning can substantially improve the model performance and lead to state-of-the-art results on both datasets. In addition, we could observe that the scores on the Clotho dataset are relatively lower than those on the AudioCaps dataset, which is consistent with the results in [16]. The reasons might be two folds: (1) the size of training data is limited in the Clotho dataset, (2) captions for each audio clip in the Clotho dataset are more diverse thus it is challenging to map the diverse captions close with each other in the shared acoustic semantic embedding space.

4.2. Metric learning objectives

As can be seen in Table 1, the NT-Xent loss shows stable performance on both AudioCaps and Clotho datasets regardless of whether the pre-trained encoders are fine-tuned or frozen. It outperforms the other three triplet-based losses in most cases on both text-to-audio and audio-to-text retrieval. For the other three triplet-based losses, the models trained via triplet-sum are not as good as others especially when the encoders are frozen. However, they achieve similar performance on the Clotho dataset when the encoders are fine-tuned. The triplet-max and triplet-weighted losses achieve similar results on the AudioCaps dataset, but the models trained via triplet-weighted loss perform less well than others on the Clotho dataset when the encoders are frozen. The weighting method does not bring substantial improvements. The reason might be that the weighting function has many hyper-parameters to be tuned while we have used the values in the original work [21]. Tuning the hyper-parameters for the dataset we evaluated may lead to better performance. Overall, the NT-Xent loss outperforms other three triplet-based losses on both text-to-audio and audio-to-text retrieval in most situations.

In our experiments, we also found that the two losses based on hardest negative mining, the triplet-max and triplet-weighted losses, are sensitive to the training settings, while the other two, triplet-sum and NT-Xent, are more robust and stable to different training settings. For example, triplet-max and triplet-weighted need a good learning rate, otherwise the models are difficult to

converge. On the AudioCaps dataset, all the models are trained with a learning rate of 1×10^{-4} and all converge well. However, the models trained via triplet-max and triplet-sum with such a learning rate do not converge on the Clotho dataset. Furthermore, triplet-max and triplet-weighted are sensitive to the initialization of model parameters, and we found that some of the models with different training seeds may not converge under the same training settings.

In addition, we study the impact of batch size on the AudioCaps dataset without fine-tuning the encoders. The results are shown in Table 2. We can observe the performance of models trained via triplet-max and triplet-weighted losses degrades considerably when the batch size is increased to 128, while models trained via the other two losses (triplet-sum and NT-Xent) show stable performance with respect to the change of the batch size. This is somewhat inconsistent with results in the literature, where it is generally believed that larger batch sizes could lead to better performance for the triplet-based losses [18]. We found the triplet-max and triplet-weighted do not converge with a batch size of 128. The reason might be that the learning rate should also be tuned to adapt for a larger batch size. Overall, triplet-sum and NT-Xent losses are more robust to different training settings and datasets, while triplet-max and triplet-weighted are more difficult to train.

5. Conclusions

We have presented a simple but effective model to learn the Acoustic Semantic Embeddings for the free-form language-based audio-text retrieval task, then we studied the impact of metric learning objectives based on our model in a constant training setting. We empirically demonstrated that the metric learning objectives have a significant impact on the model performance, where the NT-Xent loss outperformed the popular triplet-based losses and showed stable performances with respect to different training settings and datasets. The triplet losses with hard negative mining need careful tuning of hyper-parameters, and are sensitive to the initialization of model parameters.

6. Acknowledgements

7. References

[1] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. [Online]. Available: https://github.com/fartashf/vsepp

[2] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4654–4662.

[3] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[4] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *British Machine Vision Conference*, 2019.

[5] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv preprint arXiv:1804.02516*, 2018.

[6] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in *European Conference on Computer Vision (ECCV)*, 2020.

[7] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, "Large-scale content-based audio retrieval from text queries," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 105–112.

[8] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE)*, 2018, pp. 59–63.

[9] B. Elizalde, S. Zarar, and B. Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4095–4099.

[10] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 211–215.

[11] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Diverse audio captioning via adversarial training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[12] X. Liu, Q. Huang, X. Mei, T. Ko, H. Tang, M. D. Plumbley, and W. Wang, "CL4AC: A contrastive loss for audio captioning," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 196–200.

[13] X. Liu, X. Mei, Q. Huang, J. Sun, J. Zhao, H. Liu, M. D. Plumbley, V. Kılıç, and W. Wang, "Leveraging pre-trained bert for audio captioning," *arXiv preprint arXiv:2203.02838*, 2022.

[14] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 119–132.

[15] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[16] A. S. Koepke, A.-M. Oncescu, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, 2022.

[17] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *European Conference on Computer Vision*. Springer, 2020, pp. 681–699.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[19] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *Proc. Interspeech 2020*, pp. 2977–2981, 2020.

[20] M. Bleeker and M. de Rijke, "Do lessons from metric learning generalize to image-caption retrieval?" *arXiv preprint arXiv:2202.07474*, 2022.

[21] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[23] X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. ZHAO, S. Li, T. Ko, H. Tang, X. Shao, M. D. Plumbley, and W. Wang, "An encoder-decoder based audio captioning system with transfer and reinforcement learning," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 206–210.

[24] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[27] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980