3CMLF: Three-Stage Curriculum-Based Mutual Learning Framework for Audio-Text Retrieval

Yi-Wen Chao* Dongchao Yang* Rongzhi Gu* and Yuexian Zou*
* ADSPLAB, School of ECE, Peking University, Shenzhen, China
E-mail: 2001213511@pku.edu.cn

***Abstract*—Audio-text retrieval aims to retrieve instances that best match a given instance from an audio modality to a text modality and vice versa. Recent studies have mainly focused on capturing the shared high-level semantic concepts between these two modalities by synchronously updating the audio and text encoders. We found that such a synchronous updating strategy results in sub-optimal learned audio and text encoders owing to the two encoders' varying initial prior knowledge level. Furthermore, we observed a big semantic gap between the representation of audio and text encoders using the common mini-batch sampling strategy. To tackle these issues, we present a novel three-stage curriculum-based mutual learning framework (3CMLF) to boost the performance. Our approach includes two key components: (i) Inspired by the human learning process, we provide a global curriculum-based hard sample mining strategy, which can globally mine the easiest, median, and hardest negative samples from the full training set and construct three training sets respectively. (ii) We propose to train the text and audio encoders under the three-stage cross-modal mutual learning framework using the three constructed training sets. In the first stage, we fix the weights of the text network, which are initialized using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, and then update the audio encoder based on the easiest training set. During the second stage, we freeze the audio encoder and update the text network based on the median training set. After these initial alignment stages, we release all weights to be learned and fine-tuned on the hardest training set. This three-stage process is crucial for allowing the model to successfully differentiate the top retrieved instance from a hard negative set and capture the correlation between the audio-text modal. Notably, 3CMLF is adaptable to the majority of current audio-text models as it requires no alteration to the model architecture. Experimental results on the AudioCaps dataset show that our method achieves a new state-of-the-art performance.**

I. Introduction

For each given instance in text modality, audio-text retrieval task aims to retrieve the best-matching audio instances from a group of candidates and vice versa. With the vast increase in the numbers of user-generated multimedia data from online communities and application, it becomes difficult for users to effectively and efficiently search for information of interest [1]. Under such circumstances, cross-modal retrieval has attracted extensive attention in recent studies [2]–[8]. However, when compared with visual-text and other cross-modal retrieval tasks, audio-text retrieval has not received much attention in the research area of multimedia, mainly owing to a lack of appropriate datasets [9]. Therefore, early audio studies dealing with cross-modal retrieval across audio and text modalities are based on metadata, *e.g.,* an audio tag, instead of a free-form natural language query. Chechik et al. [10] addressed a system that can retrieve sounds based on single-word audio tags. To search for audio using an onomatopoeic query, Ikawa [11] measured the distance between the sound and onomatopoeic query within the shared latent spaces. Elizalde et al. [12] associated audio with text by jointly learning the audio and text representations using a twin network. Although it is viabble to retrieve metadata from manually-curated database [9], such tag-based sound retrieval frameworks have a limited performance constrained by the audio tag format. Following the publication of audio captioning datasets [13], [14], new public benchmarks were addressed by Koepke et al. [9] for audio retrieval task, using detailed free-form language as searching queries. Because natural language queries are one of the most recognized user interfaces commonly employed in existing cross-modal search engines, free-form text-based audio retrieval could contribute to a more flexible retrieval between audio and text. According to Mei et al. [15], varied metric learning objectives have considerably different effects on audio-text retrieval based on free-form natural language. The general idea behind these previous studies is to narrow down the gap in heterogeneity between audio and text modalities by synchronously learning two functions [16], *i.e.,* audio and text encoders, thereby transforming the data from a multi-modal form into a common representation space, where relevant data are closely spaced and irrelevant data are spaced widely apart.

Despite the significant advancement achieved in prior studies, there are still a number of obstacles in constructing an efficient audio-text retrieval model, which have not been properly tackled in the past research. Most methods in the prior studies synchronously update both the audio and text encoders, and such a training strategy gives little attention to the different levels of prior knowledge that audio and text encoders carry at the initialization phase. For instance, the text encoder used in our study is initialized using Bidirectional Encoder Representations from Transformers (BERT) [17], which is pre-trained from a massive set of unlabeled data and contains high-level prior knowledge. Forcing BERT and the audio encoder to be updated synchronously will result in an oscillatory optimization during the early training. In addition, we found that in the previous audio-text retrieval methods, their results in R@10, which is denoted as the percentage of correct matching

978-616-590-477-3 ©2022 APSIPA APSIPA ASC 2022

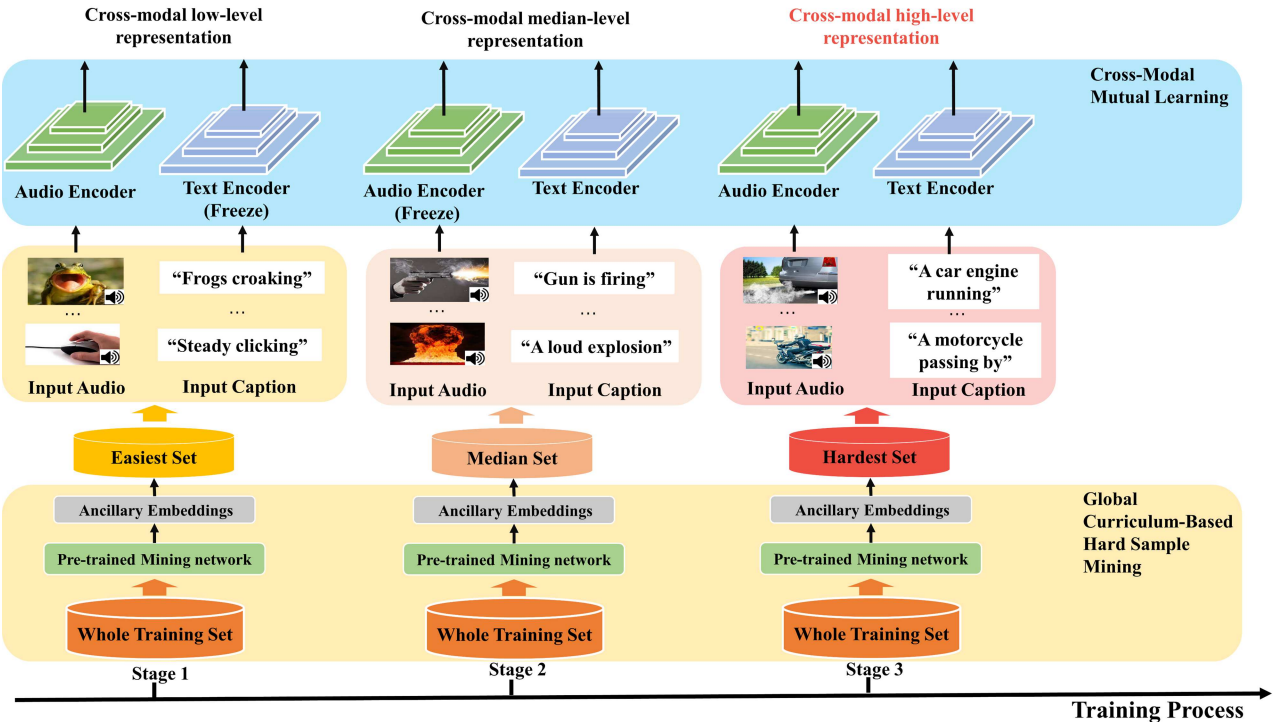

Fig. 1. Panorama of three-stage curriculum-based mutual learning framework (3CMLF), consisting of two primary parts: a global curriculum-based hard sample mining strategy and a cross-modal mutual learning framework. In stage 1, training specializes on capturing specific low-level modality features, whereas training in the later stage is meant to capture modal-invariant higher-level concepts in the representations

among the top-10 ranked retrieved results, is much higher than their results in R@1. This indicates that the correct answer is more likely to be included in the top-10 search results than in the top-1 search results. Thus, learning a fine-grained cross-modal correspondence is the key to effectively discriminating these hard samples from the correct answer. Furthermore, we observed that random sampling from the full training set might bring difficulty for the audio and text encoders to learn the proper representations. Specifically, data that share similar details of the semantic content, are complicated for model to distinguish at early training stage (*e.g.,* car engine roaring and engine starting up).

To tackle the aforementioned constraints, we propose a novel three-stage curriculum-based mutual learning framework (3CMLF) to improve the performance of the audio-text retrieval task. The proposed framework consists of two essential parts, a global curriculum-based hard sample mining strategy and a cross-modal mutual learning framework.

Specifically, as the general framework of 3CMLF illustrated in Fig. 1, we first develop a curriculum-based hard sample mining strategy. We propose a training strategy inspired by the process of humans' acquiring knowledge, which progresses from simple to more difficult samples during training. Using pre-trained ancillary embeddings computed from a pre-trained cross-modal deep embedding network [15], we calculate the semantic similarity between the training data and globally mine the easiest negative pairs to construct the first training set, *i.e., the easy training set*, where the data within each batch are semantically dissimilar. Similarly, we select the median negative pairs and the hardest negative pairs respectively to construct the second training set, *i.e., the median training set*, and the third training set, *i.e., the hard training set*. We then design three training sets to train the audio and text encoders based on the three pre-constructed training sets. In the first stage, we fix the weights of the pre-trained text encoder, which are initialized employing the pre-trained BERT model, and only update the audio encoder's parameters training on the constructed *easy training set*. In this way, the audio network can learn to align itself to the initial text representations. In the second stage, we then fix the audio encoder and only update the parameters of the text encoder based on the *median training set*. Finally, we jointly train the text and audio encoder network on the *hard training set*.

In conclusion, there are three major contributions in this article.

1) We introduce a global curriculum-based hard sample mining approach targeted to the audio-retrieval task, which can globally mine the hardest, median, and easiest samples, and accordingly construct a three-stage training set. We then explore the performance of our proposed mining strategy. The results indicate that our global curriculum-based hard sample mining strategy outperforms the local mining technique applied in the randomly sampled mini-batch.
2) We propose a cross-modal mutual learning framework

that enables the two sub-networks to learn from each other, providing extensive assistance for the model to capture the fine-grained semantic correspondence between the audio and text modals.

3) Extensive experiments are conducted on the mainstream dataset AudioCaps. The experimental result indicate that our proposed model achieves a new state-of-the-art performance.

II. Methods

A. Problem formulation

For the problem formulation of the audio-text retrieval, we assume that $D = \{d_i\}_{i=1}^{N} = \{(a_i, t_i)\}_{i=1}^{N}$ is a collection of N examples of audio-caption pairs, where a_i is the input audio clip, and t_i is the paired caption of the ith example in D. We simply consider each audio clip to have only a single paired caption. (a_i, t_i), consisting of an audio clip with its corresponding caption, is considered as a positive pair, whereas $(a_i, t_{j,j\neq i})$ is a negative pair.

Because the audio and text feature vectors typically lie within distinct representation spaces, it is not practical to make a direct comparison between them for cross-modal retrieval. [18]. Thus two encoders for audio and text modalities are learned respectively using cross-modal learning: $x_i = f(a_i; \Upsilon_a) \in \mathbb{R}^d$ and $y_i = g(t_i; \Upsilon_t) \in \mathbb{R}^d$, where f represents the audio encoder, g represents the text encoder and d stands for the dimensionality of the embedding within the joint embedding space. Υ_a and Υ_t denote the trainable parameters of the f and g. The similarity between each audio-caption pair $(\mathrm{a_i}, t_j)$ can be denoted as follows:

$$\mathrm{s}_{ij} = \frac{f(a_i; \Upsilon_a) \cdot g(t_i; \Upsilon_t)}{\|f(a_i; \Upsilon_a)\|_2 \, \|g(t_i; \Upsilon_t)\|_2} \tag{1}$$

Both of the encoders, f and g, are trained to increase the similarity of positive pairs $\mathbf{s}_{ii}$ while at the same time decrease the similarity of negative pairs $\mathbf{s}_{ij}$.

B. Global curriculum-based hard sample mining

Curriculum learning addresses the question of how to use prior knowledge regarding the difficulty of the training examples to sample each mini-batch non-uniformly and thereby increase the learning rate and accuracy. The curriculum learning paradigm is based on the premise that introducing the learner with simple concepts first helps the learning process.

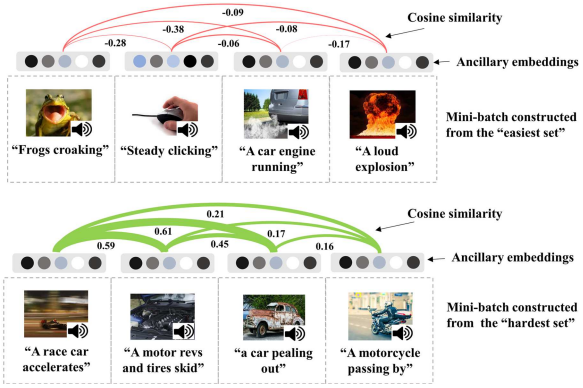

Fig. 2. Mini-batch constructed from different training sets

Following this insight, we propose a global curriculum-based hard sample mining strategy that can globally mine the easiest, median, and hardest samples and construct three training sets accordingly. Intuitively, the model should learn easier negative samples first, followed by progressively harder negative samples as the training stage proceeds and the training process converges. To be specific, we adopt pre-trained ancillary embeddings that are computed from the pre-trained cross-modal deep embedding network proposed by Mei et al. [15]. Each training sample is assigned an ancillary embedding, which is used to construct the mini-batch with suitable samples accordance to different training stages. They are vectors possessing the following properties:

1) As stated above, let $D = \{\mathrm{d_i}\}_{\mathrm{i=1}}^{N} = \{(\mathrm{a_i}, \mathrm{t_i})\}_{i=1}^{N}$ denotes the data, where a_i represents an audio clip and $\mathrm{t_i}$ represents its paired caption. Each training sample $\mathrm{d_i}$ in the dataset has a pre-trained ancillary embedding $\mathrm{e_i}$. These embeddings are employed when generating mini-batch.
2) Two *easy negative* samples' ancillary embeddings are widely apart based on the cosine similarity metric.
3) Two *hard negative* samples' ancillary embeddings are near to one another based on the cosine similarity metric.

In the different training stages (stage1, stage2, and stage3), each mini-batch is accordingly constructed:

1) During the first stage of training, the motivation is to make the data within a mini-batch as semantically dissimilar as possible. Thus, we sample a collection of mini-batches from the dataset D and attempt to minimize the objective L:

$$\mathrm{L} = \arg\min \left(\sum_{m=1}^{M} \sum_{i=1}^{B} \sum_{j\neq i, j=1}^{B} \mathrm{Cos\,Sim}(e_i, e_j) \right) \tag{2}$$

where M is the iteration and B is the batch size. As illustrated in the top of Fig. 2, in the mini-batch, the similarity scores between the samples' ancillary embeddings are rather low.
2) During the second stage of training, we randomly sample a collection of mini-batches from the dataset D.
3) During the third stage of training, the motivation is to make the data within a mini-batch as semantically similar as possible. We sample a collection of mini-batches from the dataset D and attempt to maximize the objective L. Once the mini-batch has been filled, it includes a collection of hard samples, which is crucial for the cross-modal mutual learning framework to produce a discriminative high-level representation.

$$\mathrm{L} = \arg\max \left(\sum_{m=1}^{M} \sum_{i=1}^{B} \sum_{j\neq i, j=1}^{B} \mathrm{CosiSim}(e_i, e_j) \right) \tag{3}$$

C. Audio Encoder

Pre-trained audio neural networks [19], also known as PANNs, are networks trained from AudioSet [20] (1.9 million audio clips), demonstrating state-of-the-art performance in audio tagging task. These networks demonstrate their transferability by successfully tackling six different audio pattern recognition problems. Following the prior state-of-the-art approaches in audio-text retrieval, our experiments are performed with the pre-trained ResNet-38 in PANNs, with a pooling layer replacing the last two linear layers. The pooling layer consists of (i) an average pooling along the frequency axis followed by (ii) an average and a max pooling along the time axis. The features from both pooling are summed together and fed into a simple multi-layer perceptron (MLP) block, consisting of two linear layers sandwiching a ReLU [21] activation layer. Using the MLP block, we map the audio features into a shared embedding space.

D. Text Encoder

The approach for pre-training large-scale language models based on mass unlabeled data has recently made significant strides in a number of NLP tasks [22]. BERT is pre-trained on two tasks, namely next sentence prediction and masked-language modeling, and delivers cutting-edge results for state-of-the-art results for a broad range of NLP tasks, producing a powerful contextualized word embedding. Following the prior state-of-the-art approaches, we adopt the pre-trained BERT as our text encoder in this study.

BERT takes a sequence of tokens as input, with the first token always being $[CLS]$, and returns the last hidden state of $[CLS]$ as the entire sequence's representation. To map the text representation into the shared embedding space, an MLP block is also employed.

E. Cross-Modal Mutual Learning framework

As stated previously, we found that if we train both models simultaneously, which means forcing BERT and the audio encoder to be updated synchronously will result in an oscillatory optimization during the early training. To this end, we design an effective cross-modal mutual learning framework to transfer the fine-grained semantic knowledge between the two encoders that have different levels of prior knowledge.

Our cross-modal mutual learning framework consists of three stages. Specifically, in the first stage, the text encoder, which is initialized using a pre-trained BERT, has higher levels of prior knowledge than the audio encoder. Thus we fix the parameters of the text encoder (teacher model) , and update only the parameters of the audio encoder (student model) based on the easy training set. This way the audio encoder can learn to align itself to the initial text representation and reach a higher knowledge level than text encoder. In the second stage, we freeze the audio encoder (teacher model) and update the text encoder (student model) in the same way based on the median training set. After these initial alignment stages, we release all weights to be learned and trained on the hardest training set. The mutual learning between the two sub-network in stage1 and stage 2 will further boost the jointly updating process in stage 3. This three-stage process is crucial for allowing the model to successfully capture the fine-grained correspondence between the audio modal and the text modal, leading to better model performance.

F. Loss Function

The normalized temperature-scale cross-entropy (NT-Xent) loss [23] is a commonly used softmax-based loss function for contrastive representation learning. In a previous study on the audio-text retrieval task, the NT-Xent loss achieved better performance than the commonly used triplet-based losses [24], [25] and was more robust to different training settings [15]. Thus, in this study, the proposed 3CMLF is trained using the NT-Xent loss, which is defined as:

$$\mathcal{L} = -\frac{1}{B}\left(\sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} + \sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}/\tau)}\right), \tag{4}$$

where B is the batch size, and τ is the temperature hyper-parameter. The NT-Xent loss function in this case contains two term since the audio-text retrieval task includes both audio-to-text retrieval and vice versa. The target of NT-Xent is to maximize positive pair's similarity with reagard to all negative pairs in a mini-batch, bidirectionally.

III. Experiments and Analysis

A. Dataset

In our research, We use AudioCaps dataset, which contains about 49274 audio clips in the training set. There are 494 and 957 audio clips in the validation and test sets, respectively. All audio samples in the AudioCaps dataset are approximately 10s long. Each audio is human-annotated with a single reference caption in the training set and five reference captions in the validation and test sets.

B. Implementation Details

We extracted 64 dimensional Log mel-spectrograms, employing a 1024-points Hanning window with a 320-points window hop size, as the input features. The proposed 3CMLF is trained with batches of 32 for at most 50 epochs using the Adam optimizer [26]. The learning rate is set to 1 × 1e-4 and is decreased by 1/10th every 20 epochs. Following the settings employed in the prior study, for NT-Xent, the temperature hyper-parameter $\tau = 0.07$. We set the dimension of the joint embedding space to 1024. All tests are conducted on the RTX3090 GPU.

C. Evaluation Metrics

The audio-text reatrieval performance is measured in terms of recall at rank k ($R@k$), which is a commonly used cross-modal retrieval evaluation metric. $R@k$ is denoted as the percentage of correct matching within the top-k ranked results. We report the results for $R@1$, $R@5$, and $R@10$.

TABLE I
COMPARISON BETWEEN DIFFERENT TRAINING STRATEGIES FOR 3CMLF AND OTHER PREVIOUS STATE-OF-THE-ART METHODS. CHM DENOTES THE CURRICULUM-BASED HARD SAMPLE MINING. AUDIO ENCODER DENOTES THE TRAINING STRATEGY FOR THE AUDIO ENCODER

Model	**Audio encoder**	**With CHM**	**Text-to-Audio**			**Audio-to-Text**		
			$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
CE [9]			23.6	56.2	71.4	27.6	60.5	74.7
MoEE [9]			23.0	55.7	71.0	26.6	59.3	73.5
ResNet38+BERT [15]	from scratch		24.5	56.9	71.6	30.8	59.8	75.8
	pre-trained		33.7	69.5	82.4	38.7	71.6	83.8
3CMLF	from scratch	No	26.5	60.1	74.5	31.9	63.8	76.6
	from scratch	Yes	28.0	60.5	75.1	33.0	65.0	78.1
	pre-trained	No	34.5	69.9	82.3	40.5	72.0	84.8
	pre-trained	Yes	**34.9**	**70.7**	**82.9**	**41.4**	**72.9**	**85.4**

TABLE II
ABLATION STUDY OF OUR PROPOSED 3CMLF AT DIFFERENT TRAINING STAGES

Model	**Audio encoder**	**Training stage**	**Text-to-Audio**			**Audio-to-Text**		
			$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
3CMLF	from scratch	stage 1	9.4	30.1	43.1	9.8	29.8	44.0
	from scratch	stage 2	19.3	49.2	65.3	24.1	50.9	64.6
	from scratch	stage 3	28.0	60.5	75.1	33.0	65.0	78.1
	pre-trained	stage 1	18.1	48.9	65.2	21.5	51.3	67.2
	pre-trained	stage 2	29.2	65.4	79.9	35.1	67.1	79.6
	pre-trained	stage 3	**34.9**	**70.7**	**82.9**	**41.4**	**72.9**	**85.4**

TABLE III
EXPERIMENTAL RESULTS WITH DIFFERENT BATCH SIZES.

Model	**Audio encoder**	**Batch size**	**Text-to-Audio**			**Audio-to-Text**		
			$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
3CMLF	from scratch	16	24.2	57.3	72.6	29.7	60.6	74.3
	from scratch	32	28.0	60.5	75.1	33.0	65.0	78.1
	from scratch	64	27.1	59.1	73.5	32.0	64.6	78.2
	pre-trained	16	28.2	64.9	80.9	35.4	64.9	80.2
	pre-trained	32	**34.9**	**70.7**	**82.9**	**41.4**	**72.9**	**85.4**
	pre-trained	64	33.1	68.3	81.8	38.7	70.5	83.1

D. Model Performance

Table I presents the detailed results of our experiment. It can be demonstrated that the global curriculum-based hard sample mining strategy can significantly enhance the model performance and provide state-of-the-art outcomes. As Table I shows, the proposed 3CMLF achieves a better result than the baseline regardless of whether the audio encoder is pre-trained or trained from scratch. In addition, it outperforms three prior state-of-the-art approaches on text-to-audio task and vice versa. Koepke et al. [9] addressed new benchmarks for audio-text retrieval task. They adopted robust cross-modal video retrieval approaches to audio-text retrieval task, including MoEE and CE, and provided the baseline results. Following the benchmark, the baseline model [15] used in our study explored the impact of different metric learning objectives, leading to state-of-the-art result on the AudioCaps dataset. Reproduced experimental results are employed as the baseline model's performance.

E. Ablation study on different training stages

The application of global curriculum-based hard sample mining to a cross-modal mutual learning framework is based on the assumption that the audio and text encoders enjoy different levels of prior knowledge at the initialization phase. Thus, when both encoders are synchronously updated, it may result in a sub-optimal training process. To prove this hypothesis, we studied the model performance for different training stages, and explored the influence of within-modality self-instance discrimination and cross-modal discrimination. The results are reported in Table II. In Table II, we can see that when the transfer process is completed after stage 3, it results in better representations than stage 1 and stage 2 owing to the curriculum-based knowledge transfer between audio and text. From these results, we can conclude that this three-stage process is crucial for the model to successfully capture a fine-grained high-level correspondence between audio-text modals. In addition, the mutual learning framework, which enables two sub-networks from different modalities to learn from each other, effectively produces a better model.

F. Effects of different batch sizes

Further, we investigated how the model performance was affected by different batch sizes. Table III shows the performances of the models based on different training strategies applied to the audio encoder. The performance of 3CMLF is quite stable when the batch size is increased to 64. The model

performance on audio-to-text retrieval and text-to-audio retrieval degrades considerably when the batch size is decreased to 16. In addition, the strategy for constructing a mini-batch should be modified to adapt to a different batch size.

IV. Conclusions

In this paper,we introduced an efficient model to capture the high-level semantic correspondence between the audio and text modals. We proved experimentally that the global curriculum-based hard sample mining strategy and the cross-modal mutual learning framework had a substantial effect on the performance of the natural language based audio-text retrieval, in which 3CMLF outperformed the prior state-of-the-art methods and demonstrated a steady performance with regard to different training strategies and settings, leading to a new state-of-the-art results.

Acknowledgment

This paper was partially supported by NSFC (No: 62176008), Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001)

References

[1] Q. Feng, P. Li, Z. Lu, G. Liu, and F. Huang, "End-to-end learning for encrypted image retrieval," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1839–1845.

[2] J. Dong, X. Li, and C. G. Snoek, "Word2visualvec: Image and video to sentence matching by visual feature prediction," *arXiv preprint arXiv:1604.06838*, 2016.

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

[4] 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 *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 19–27.

[5] 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.

[6] 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.

[7] J. Wehrmann, C. K. dos Reis, and R. C. Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020, Brasil.*, 2020, pp. 12 313–12 320.

[8] Y. Gong, G. Cosma, and H. Fang, "On the limitations of visual-semantic embedding networks for image-to-text information retrieval," *Journal of Imaging*, vol. 7, no. 8, p. 125, 2021.

[9] 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.

[10] 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.

[11] S. Ikawa and K. Kashino, "Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds." in *DCASE*, 2018, pp. 59–63.

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

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

[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, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[15] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "On metric learning for audio-text cross-modal retrieval," *arXiv preprint arXiv:2203.15537*, 2022.

[16] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8867–8871.

[17] 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*, vol. 1(Long and Short Papers), 2019, pp. 4171–4186.

[18] W. T. Tseng, C. Y. Wu, Y. C. Hsu, and B. Chen, "Faq retrieval using question-aware graph convolutional network and contextualized language model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 2006–2012.

[19] 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.

[20] J. F. Gemmeke, D. P. 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[21] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 11–13 Apr 2011, pp. 315–323.

[22] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[23] 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.

[24] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improved visual-semantic embeddings," *arXiv preprint arXiv:1707.05612*, 2017.

[25] 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*, 2020, pp. 13 005–13 014.

[26] B. Jimmy and P. Diederik, "Adam: A method for stochastic optimization," *arXiv preprint arXiv: 1412.6980*, 2014.