LANGUAGE-BASED AUDIO RETRIEVAL WITH PRE-TRAINED MODELS

Technical Report

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

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

ABSTRACT

This technical report presents a language-based audio retrieval system that we submitted to Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2022 Task 6b. Language-based audio retrieval is a cross-modal task aiming at retrieving a matched audio clip from a pool of candidates given a language query such as a sentence. Cross-modal retrieval tasks are often solved by using deep learning models where the features from different modalities are extracted and then mapped to a joint embedding space. These models usually require a large amount of training data to obtain reasonable performance. However, the audio captioning dataset employed in this audio retrieval task is limited in size. In this work, we propose to use large-scale pre-trained models as both audio and text encoders to mitigate the data scarcity problem and learn the acoustic semantic embeddings. Results on the Clotho dataset show that our proposed system significantly improves the scores of all the evaluation metrics as compared to the baseline system.

Index Terms— Cross-modal task, audio retrieval, deep learning, text-based retrieval

1. INTRODUCTION

Language-based audio retrieval aims at retrieving a matched audio clip from a pool of candidates using a natural language query such as a sentence [1], which can be applied in several applications such as multimedia data retrieval, audio book production and web search. This task has received limited attention in the literature due to the lack of appropriate datasets [1]. With the release of audio captioning datasets [2, 3], Koepke et al. [1] first investigated and established three public benchmarks for this task. This task has been included in DCASE Challenge 2022 [4] to encourage further research in this area, and this technical report describes our audio retrieval system submitted to DCASE Challenge Task 6b.

Cross-modal retrieval tasks are generally solved by using deep learning models to extract features from different modalities, and then map the extracted features to a joint embedding space for similarity comparison [5, 6, 7, 8]. Our proposed system consists of an audio encoder and a text encoder, which extracts the features for audio and text respectively. Because the availability of data is limited in existing audio captioning datasets, we make use of pre-trained models to learn robust features via transfer learning. Specifically, pre-trained audio neural networks (PANNs) [9] pre-trained on AudioSet [10] are employed as the audio encoders while BERT [11] is used as the text encoder. The proposed model is trained via the normalized temperature-scaled cross entropy loss (NT-Xent) [12] due to its stable performance compared with other triplet-based losses [13]. Ensemble is used to further improve the system performance. Results on the Clotho dataset show that our proposed system significantly improves the scores of all the evaluation metrics compared to the baseline system [14].

The remainder of this technical report is organized as follows. In Section 2, we introduce the proposed system in detail. The experiments and results are discussed in Section 3. Finally, this work is concluded in Section 4.

2. SYSTEM DESCRIPTION

In this section, we first formulate the language-based audio retrieval problem, and then introduce our proposed system.

2.1. Problem Formulation

Assume we have an audio captioning dataset $D = \{(a_i, t_i)\}_{i=1}^{N}$ with N examples, where a_i is an audio clip and t_i is the paired caption. For simplicity, we assume that each audio clip just has a single corresponding caption. An audio clip with its paired caption (a_i, t_i) can be regarded as a positive pair, while $(a_i, t_{j,j\neq i})$ and $(a_{j,j\neq i}, t_i)$ are negative pairs. A good audio retrieval system should return the paired audio clip when a caption is received as the input query.

Similar to other cross-modal retrieval models [6, 5, 8], an audio encoder $f(\cdot)$ and a text encoder $g(\cdot)$ are employed to extract the audio and text features, which are then projected into a shared embedding space. Therefore, the similarity of an audio clip and a caption can be measured by e.g. cosine similarity of their embeddings in the shared embedding space. For example, the similarity of an audio-caption pair (a_i, t_j) can be defined as:

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

The two encoders are trained to increase the similarity scores of positive pairs s_{ii} while decreasing the similarity scores for negative pairs $s_{ij,i\neq j}$. During inference, the retrieval system calculates the text embedding for the queried caption using the text encoder $g(\cdot)$, which is then compared with all the audio embeddings in the dataset computed by the audio encoder $f(\cdot)$. The audio clip that gives highest cosine similarity score will be retrieved.

2.2. Model Architecture

To address the data scarcity problem and learn robust feature representation, pre-trained models are employed here.

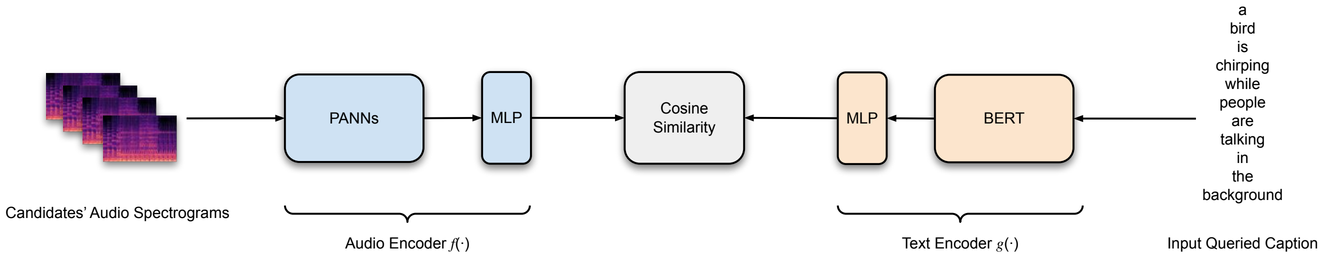

Figure 1: Overview of the language-based audio retrieval system which consists of an audio encoder and a text encoder. The cosine similarity scores are computed between the text embedding of input queried caption encoded by the text encoder $g(\cdot)$ and audio embeddings of audio candidates in the dataset encoded by the audio encoder $f(\cdot)$. The audio clip with highest cosine similarity score is finally retrieved.

Table 1: Results on the Clotho development-testing set compared with official baseline. AC: the model is first pre-trained on AudioCaps then fine-tuned on Clotho.

Model	**R**@1	**R**@5	**R**@10	**mAP**@10
Baseline [14]	0.03	0.11	0.19	0.07
CNN14+BERT	0.147	0.377	0.495	0.244
ResNet38+BERT	0.143	0.369	0.491	0.239
CNN14+BERT+AC	0.143	0.374	0.498	0.243

2.2.1. Audio Encoder

We perform experiments with two networks from PANNs [9], namely, ResNet38 and CNN14. The last two linear layers after the convolutional blocks are replaced by a new mluti-layer perceptron (MLP) block. The MLP block contains two linear layers with a ReLU [15] activation layer between them. After getting the feature map from the last convolutional block, an average pooling is first applied along the frequency dimension, and then maximum and average operations are used along the time dimension. We sum the maximized and averaged features and then project them into the shared embedding space through the MLP block.

2.2.2. Text Encoder

Pre-trained BERT [11], which stands for Bidirectional Encoder Representations from Transformers, is used as the text encoder due to its powerful ability at extracting contextualized word representations. To get a representation for a sentence, a "<CLS>" token is appended at the start of the sentence, and the output of that token is used as the final sentence representation. A multi-layer perceptron (MLP) is also applied to project the sentence representation to the shared embedding space. The MLP blocks in the audio and text encoders are independent.

3. EXPERIMENTS

3.1. Datasets

Clotho Clotho v2 [3] is used as the official ranking dataset of DCASE Challenge 2022 Task 6b. The published development set contains 5929 audio clips, each of which has five reference captions. For our submitted systems, 100 audio clips are randomly selected for validation and the remaining 5829 audio clips are used for training the models.

The published development set consists of three sub-sets. For comparison with other works, we report the results on the development-testing set containing 1045 audio clips, and the models are trained and validated using the development-training and development-validation sets, respectively, each containing 3839 and 1045 audio clips.

AudioCaps Koepke et al. [1] found pre-training their model on the largest audio captioning dataset, AudioCaps [2], can improve the system performance on the Clotho dataset. Therefore, we also investigate pre-training our model on AudioCaps.

All the audio clips in AudioCaps are 10-seconds long and are sourced from AudioSet [10]. In our downloaded version of AudioCaps, there 49 274 audio clips in the training set and each audio clip has one human-annotated caption. The validation and test sets have 494 and 957 audio clips, respectively, and each audio clip has five reference captions.

3.2. Experimental Setups

All the captions in the dataset are converted to lower case with punctuation removed. Log mel-spectrograms are used as the acoustic features, which are extracted using a Hanning window of 1024-points with a hop size of 320-points and 64 mel bins. All the models are trained with Adam optimizer [16] for 50 epochs. The batch size is 32 and the learning rate is set to 1×10^{-4} that is decayed by a factor of 10 every 20 epochs. SpecAugment [17] is applied during training. The dimension of the shared embedding space is 1024.

The model is trained with the NT-Xent [12] loss, because it performs more stable than other popular triplet-based losses [13]. The NT-Xent loss can be formulated as:

$$\mathcal{L} = -\frac{1}{B}\left(\sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp\left(s_{ij}/\tau\right)} + \sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp\left(s_{ji}/\tau\right)}\right), \tag{2}$$

where τ is a temperature hyper-parameter set to 0.07 following the setting in [12] and B is the batch size.

3.3. Results

The results are shown in Table 1. It can be seen clearly that our proposed methods significantly outperform the baseline on all the metrics. The two audio encoders, ResNet38 and CNN14, achieve similar performance. It is interesting to note that pre-training our model on AudioCaps does not improve the system performance on the Clotho dataset. The reason might be that the pre-trained models in our system are fine-tuned, while this is not the case for Koepke et al. [1]. As a result, pre-training the model showed performance improvements in their case.

Our submitted systems to DCASE Task 6b are summarized as follows:

- Submission 1: Single model with CNN14 audio encoder.
- Submission 2: Ensemble of models with CNN14 audio encoder.
- Submission 3: Ensemble of models with ResNet38 audio encoder.
- Submission 4: Ensemble of models pre-trained on AudioCaps with CNN14 audio encoder.

4. CONCLUSION

This technical report has briefly described our system submitted to DCASE 2022 Task 6b. We make use of pre-trained models, PANNs and BERT, as audio encoder and text encoder, and fine-tune them on the Clotho dataset to address the data scarcity problem. The results show our method significantly improves all the metrics as compared with the baseline system.

5. ACKNOWLEDGMENT

This work is partly supported by a Newton Institutional Links Award from the British Council, titled "Automated Captioning of Image and Audio for Visually and Hearing Impaired" (Grant number 623805725) and a grant EP/T019751/1 from the Engineering and Physical Sciences Research Council (EPSRC). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] http://dcase.community/challenge2022/.

[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] 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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680