LANGUAGE-BASED AUDIO RETRIEVAL WITH PRETRAINED CNN AND GRAPH ATTENTION

Technical Report

Feiyang Xiao[1], *Jian Guan*[1*], *Haiyan Lan*[1], *Qiaoxi Zhu*[2], *and Wenwu Wang*[3]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Centre for Audio, Acoustic and Vibration, University of Technology Sydney, Ultimo, Australia
[3]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

ABSTRACT

This technical report describes our system submitted for Task 6B of the DCASE2022 Challenge (language-based audio retrieval). Our audio retrieval system has an audio encoder composed of a pretrained CNN module (i.e., pretrained audio neural network, PANNs) and a novel graph attention module. Its text encoder is the pretrained word2vec model, which is the same as in the baseline system of Task 6B. Experiments show that our audio retrieval system can achieve an mAP10 metric (used for ranking in the DCASE Challenge) of 13% on the development-testing dataset of Task 6B.

***Index Terms*—** Audio retrieval, audio neural network, graph attention.

1. INTRODUCTION

Language-based audio retrieval aims to retrieve audio signals using textual descriptions (i.e., audio captions) of the sound content [1]. It is a language-based retrieval different from the previous audio retrieval research, which searches for audio signals that match an audio query [2, 3]. Furthermore, the language-based audio retrieval benefits the natural language queries widely used in current search engines [4]. Our language-based audio retrieval system uses an audio encoder of a pretrained audio neural network (PANNs) module [5] and a graph attention [6] based module to extract effective audio features [7]. Its text encoder employs the pretrained word2vec language module. As a result, our system achieves 13% in mAP10 metric on the development-testing data of DCASE2022 Challenge Task 6B.

2. SYSTEM DESCRIPTION

The structure of our audio retrieval system is shown in Figure 1. It has two inputs, namely, audio signals and the caption query. The audio signals are processed as the log-Mel spectrograms and the audio embedding is generated as the output through the PANNs encoder [5] and the graph attention (GAT) module. The caption query is processed through the text encoder using a pretrained word2vec model which generates the word embedding as outputs. Compared with the baseline system [1], we replace the CRNN module with the PANNs module [5] together with our proposed graph attention module.

In the audio encoder, the PANNs module (i.e., CNN10) includes four convolution blocks and two linear layers. Each convolution block has two convolution layers with a kernel size of 3×3 and an average pooling layer. We use the global average pooling on the Mel-band dimension of the PANNs module's output to reduce the computation cost. In the text encoder, the word embeddings of the input caption query are squeezed on the word sequence dimension to reduce the computation cost.

3. EXPERIMENTS

3.1. Data Processing

We use the development dataset of DCASE2022 Challenge Task 6B to train our system. Specifically, the development, validation, and evaluation split in the Clotho-v2 dataset [8] are used, respectively, as the development-training, development-validation and development-testing data for DCASE2022 Challenge Task 6B.

In the experiment, the raw audio signals have a sample rate of 44.1 kHz, and their log-Mel spectrograms are used as the inputs for the audio encoder. Specifically, we use 64 log-Mel band spectrogram, obtained by applying Hamming window with 50% overlap.

3.2. Setup

The input word embedding used in our ensemble system is from a word2vec language model [9] that is pretrained on the captions from the development dataset of Task 6B. The feature dimension of the audio embedding and the word embedding are both set as 300. The system is optimized by the triplet loss function with the Adam optimizer [10] and the learning rate is set as 0.0001. The batch size is set as 16 in the training process.

3.3. Performance Evaluation

Following the rules of Task 6B, we evaluate the performance with the recall and mAP metrics. Specifically, we evaluate the recall of the most relevant 1, 5 and 10 searched results, and the mAP of the most relevant 10 results.

The performance comparison of our system and the baseline system is shown in Table 1. Our system outperforms the baseline system in all the evaluation metrics, delivering better language-

Corresponding author.

This work was supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010.

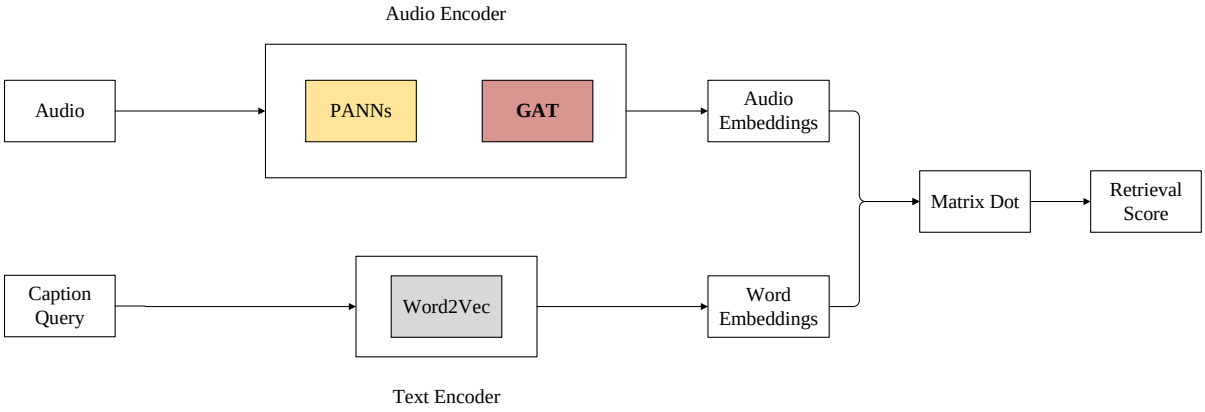

Figure 1: The structure of our proposed audio retrieval system, with audio encoder of pretrained audio neural network (PANNs) and graph attention (GAT).

Table 1: Performance results on the development-testing dataset of Task 6B.

Method	R1(%)	R5(%)	R10(%)	mAP10(%)
Baseline system [1]	3	11	19	7
Our system	**7**	**21**	**33**	**13**

based audio retrieval performance using the pre-trained CNN module and graph attention strategy.

4. CONCLUSION

In this technical report, we have presented our system for Task 6B of the DCASE2022 Challenge. Our system uses the pre-trained CNN and graph attention in the audio encoder and achieves 13% in mAP10 metric tested on the development-testing dataset of DCASE2022 Challenge Task 6B.

5. 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 *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8867–8871.

[2] P. Manocha, R. Badlani, A. Kumar, A. Shah, B. Elizalde, and B. Raj, "Content-based representations of audio using siamese neural networks," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3136–3140.

[3] I. Lallemand, D. Schwarz, and T. Artières, "Content-based retrieval of environmental sounds by multiresolution analysis," in *Proceedings of International Conference on Systems, Man, and Cybernetics (SMC)*, 2012.

[4] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," *arXiv preprint arXiv:2105.02192*, 2021.

[5] 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.

[6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[7] F. Xiao, J. Guan, Q. Zhu, and W. Wang, "Graph attention for automated audio captioning," *IEEE Signal Processing Letters*, 2022 (submitted).

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

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of International Conference on Learning Representations (ICLR).*, 2013.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.