DCASE 2023 TASK 6B: TEXT-TO-AUDIO RETRIEVAL USING PRETRAINED MODELS

Technical Report

Chung-Che Wang[*]*, Jiawei Du*[*]*, Jyh-Shing Roger Jang*

Dept. of CS, National Taiwan University, Taipei, Taiwan
geniusturtle6174@gmail.com, {jiawei.du, jang}@mirlab.org

ABSTRACT

This technical report describes our methods to Task 6b of the DCASE 2023 challenge: Language-Based Audio Retrieval. In this work, we use the bi-encoder structure and investigate the effectiveness of different pretrained audio and text encoders, including CNN14 of PANNs, Audio spectrogram transformer, and BERT. We also try to use random deletion as data augmentation for text data, and multi-label classification as an auxiliary task for audio data.

Index Terms— Pre-trained encoders, random deletion, multi-label classification

1. INTRODUCTION

The goal of Task 6b of the DCASE 2023 challenge is to retrieve audio clips using a given caption, which is a cross-modal retrieval problem. Most of the recent methods [1, 2] use the bi-encoder structure, where different encoders are used for different modalities. In this work, we investigate the effectiveness of different pretrained audio and text encoders, including CNN14 of PANNs, Audio spectrogram transformer, and BERT. We also try to add auxiliary task and use different data augmentation schemes.

The rest of this report is organized as follows. Section 2 describes details of our approaches and submissions. Section 3 shows experimental results.

2. OUR APPROACHES AND SUBMISSIONS

Different pretrained encoders and fine-tuning schemes are tried in different submitted systems. All submissions are fine-tuned using the Clotho-development and the Clotho-validation set, and evaluated using the Clotho-evaluation set [3]. Details of differences are described in the following subsections.

2.1. System 1 and 2

System 1 is similar to the work of Mei et al. [2], where the CNN14 of PANNs [4] is used as the audio encoder, and the pretrained BERT [5] is used as the text encoder. A Multi-label classification task, which shows improvement in preliminary experiments, is used as an auxiliary task for the audio side at the training stage, where the labels are originated from the keywords provided in the metadata. For data augmentation, specAugment [6] is used for audio input, and random deletion of at most 1 token is applied for text input. The dimension of the output embedding vector is 256. The effectiveness of cross attention [7, 1] is also investigated. A 1-layer transformer with 4 heads of attention is used. The axis of attention for audio data is the frequency axis, which is determined based on preliminary experiments.

The pretrained models are fine-tuned on the Clotho dataset [3]. The loss function of this system is the sum of the NT-Xent loss [8] for embedding vectors and the cross-entropy loss for the classification task, where the cosine similarity matrix for the NT-Xent loss is the weighted sum of the cosine similarity matrix calculated by only encoders' output and the cosine similarity matrix calculated by the cross attention output.

System 2 is the same as system 1, but the cross attention is not applied. Moreover, to save computation time, both systems use only encoders' output at the inference stage (i.e. cross attention is not applied at the inference stage).

System	R@1	R@5	R@10	mAP@10
1	0.153	0.407	0.544	0.260
2	0.167	0.410	0.539	0.271
3	0.186	0.452	0.585	0.314
4	0.191	0.441	0.581	0.313

Table 1: The evaluation results on the Clotho-evaluation dataset for our submitted systems.

2.2. System 3 and 4

VALOR [9] is used for system 3 and 4, where the audio encoder is the audio spectrogram transformer [10], and the text encoder is a pretrained BERT. Despite that the Clotho dataset is included when training the VALOR models [9], we still fine-tune the models only on the Clotho dataset [3] to achieve better results. Besides, the maximal lengths of an input sentence are respectively set to 50 and 60 for system 3 and 4. For other settings including the loss functions, the default ones are kept.

3. EXPERIMENTAL RESULTS

The evaluation results on the Clotho-evaluation dataset are shown in Table 1. These results suggest that cross attention may not be suitable for text-to-audio retrieval problem, and pretrained model using large datasets is helpful.

4. ACKNOWLEDGMENT

We thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

[*]These authors contributed equally to this work

5. REFERENCES

[1] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[2] X. Mei, X. Liu, H. Liu, J. Sun, and W. W. Mark D. Plumbley and, "Language-based audio retrieval with pre-trained models," Tech. Rep., 2022.

[3] 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.

[4] 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.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[7] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 201–216.

[8] 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.

[9] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *arXiv preprint arXiv:2304.08345*, 2023.

[10] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.