ENSEMBLE SYSTEMS WITH PRETRAINED DUAL-ENCODERS FOR LANGUAGE-BASED AUDIO RETRIEVAL

Technical Report

Jiafeng Li, Xichang Cai, Shenghao Liu, Liangxiao Zuo, Menglong Wu*

North China University of Technology, Beijing, China
caixc_ip@126.com

ABSTRACT

This article presents our system developed for Task 8 of the DCASE2024 Challenge, which focuses on audio retrieval using natural language queries. Our submission incorporates a retrieval system that integrates a frozen pre-trained audio encoder and RoBERT as a text encoder. We adopted a methodology similar to the CLAP framework, training our model using paired data from the AudioCaps and Clotho datasets. Our best-performing system achieved a mean Average Precision (mAP) of 29.6% and a Recall at 1 (R@1) of 18.6% on the Clotho evaluation set.

***Index Terms*—** language-based audio retrieval, pre-trained audio encoders, RoBERT

1. INTRODUCTION

DCASE 2024 Task 8 focuses on audio retrieval through natural language [1], which is a significant problem in the field of cross-modal research. This subtask is dedicated to retrieving audio signals by utilizing textual descriptions of their sound content, known as audio captions. Human-written audio captions serve as text queries for the retrieval task. The main goal for each text query is to precisely select 10 audio files from a large dataset based on their relevance to the query, and subsequently rank them according to how well they match the query.

In this paper, we mainly employs two methods to improve the performance of the baseline system. The first method involves using data augmentation to expand the dataset, thereby increasing the size of the original training data and providing more samples for model training. The second method involves using higher-performing pre-trained audio and text encoders to extract richer features, thus enhancing the model's ability to understand audio and text information.

In this paper, the details of the system are presented in Section 2, the experimental setup and the results of the experiments are presented in Section 3, the conclusions is in Section 4.

2. METHOD

In the field of audio-related multimodal research, the CLAP [2] model architecture is simple yet highly effective. Therefore, we adopted a similar dual-encoder architecture inspired by CLAP, which includes an audio encoder and a text encoder. This architecture was trained on the AudioCaps and Clotho datasets.

* Corresponding author.

2.1. Data

We used the AudioCaps dataset, the original development set of Clotho, and the data-augmented development set of Clotho together as training data to train the model. To accurately evaluate the model's performance, we employed the evaluation set of Clotho as the test set.

AudioCaps[3] is a large-scale dataset consisting of approximately 50,000 pairs of 10-second audio clips and manually written texts. During the model training process, we only utilized the training set from this dataset. The Clotho [4] dataset contains 4981 audio samples and a total of 24905 captions. Specifically, the development split includes 2893 audio samples and 14465 captions, the evaluation split includes 1045 audio samples and 5225 captions, and the testing split includes 1043 audio samples and 5215 captions. Each audio file has five captions, providing the dataset with a rich diversity of captions.

By applying the Mixup method for audio data augmentation and using Chat-GPT to generate new captions for the mixed audio, we ultimately obtained 50K pairs of audio-text data[5]. This approach not only increases the quantity of training data but also enhances its complexity and diversity, thereby improving the model's performance.

2.2. Model

In this work, we designed two systems and integrated them using a model ensemble approach to create the final audio retrieval system for submission. Both models utilize RoBERTa [6] as an efficient text encoder, with their core difference being the use of different types of audio encoders.

The Audio Encoder: We used BEATs [7]and CNN14 [8] as our audio feature extractors. BEATs is an iterative audio pre-training framework and a self-supervised model for audio representation learning. It has shown great performance in tasks like audio classification and can generalize well to downstream tasks. We utilized the frozen parameters of BEATs as our feature extractor, resulting in a time-contextualized embedding sequence of dimension 768 for each audio. We employed CNN14 in the same manner as the baseline system.

The Text Encoder: In our model, RoBERTa is used as the text encoder. RoBERTa is an improved version of BERT, trained on a larger dataset for a longer duration, and has exhibited excellent performance across various natural language pro-

cessing tasks. It can capture rich contextual information, which helps in generating high-quality sentence embeddings.

In summary, we designed two systems, which are:

1. CNN14-RoBERTa: This system uses CNN14 as the audio encoder and RoBERTa as the text encoder.

2. BEATs-RoBERTa: This system uses BEATs as the audio encoder and RoBERTa as the text encoder.

Ensemble Methods: By employing a weighted fusion method on the two systems mentioned above, we ultimately obtained our final submission for the retrieval system. We submitted two audio retrieval systems, with their main difference being the weights assigned to each system. The two ensemble systems for submissions as follows:

1. Submission 1: The weights of CNN14-RoBERTa and BEATs-RoBERTa are both set to 0.5.

2. Submission 2: The weight of CNN14-RoBERTa is 0.32, and the weight of BEATs-RoBERTa is 0.68.

3. EXPERIMENT

During the training process, we utilized 80 epochs with a batch size of 32 and an initial learning rate of 0.001. We employed the Adam optimizer and used the same loss function as the baseline. The evaluation results on the Clotho-evaluation dataset are presented in Table 1.

Method	R1	R5	R10	mAP10
CNN14-RoBERTa	14.58	38.70	51.73	24.80
BEATs-RoBERTa	16.63	42.99	55.87	27.79
Submission 1	18.01	43.98	57.63	29.23
Submission 2	18.58	44.29	57.72	29.64

Table 1: Results on Clotho evaluation set

4. CONCLUSIONS

This article introduces the system we submitted for Task 8 of DCASE2024. The system utilizes pre-trained audio encoders and text encoders to implement an audio-based text retrieval system. Through training on AudioCaps and Clotho, the system achieved better performance than the baseline.

5. REFERENCES

[1] http://dcase.community/workshop2024/.

[2] B. Elizalde, S. Deshmukh, M. A. Ismail and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5.

[3] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. of the North American Ch. of the Ass. for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019.

[4] 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.

[5] S. -L. Wu et al., "Improving Audio Captioning Models with Fine-Grained Audio Features, Text Embedding Supervision, and LLM Mix-Up Augmentation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 316-320.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[7] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," Dec. 2022.

[8] 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," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880-2894, 2020.