OVERCOMING DATA SHORTAGE IN AUDIO-TEXT MULTI-MODAL RETRIEVAL: A TECH REPORT FOR DCASE 2023 CHALLENGE

Technical Report

Jinhee Kim[1], *Chang-Bin Jeon*[1], *Yoori Oh*[1], *JoonHyeon Bae*[2], *Kyogu Lee*[1,2,3]

[1]Department of Intelligence and Information, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]AI Institute, Seoul National University, Seoul, Republic of Korea
{ginnykim9, vinyne, yoori0203, outersky, kglee}@snu.ac.kr

ABSTRACT

This technical report proposes an audio-text retrieval model for DCASE 2023 language-based audio retrieval challenge. We focus to overcome the shortage of data in this task. To this end, we propose two approaches: the first involves gathering large paired audio-text datasets, while the second employs various augmentation techniques such as PairMix and Multi-TTA. Our experimental evaluations demonstrate the effectiveness of these approaches, while achieving competitive performance in audio-text multi-modal retrieval tasks.

***Index Terms*—** Data augmentation, Audio-language learning, Audio-text retrieval

1. INTRODUCTION

The 2023 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2023) Task 6, automated audio captioning and language-based audio retrieval challenge, presents a unique opportunity to explore the retrieval of audio content based on associated textual information. However, one of the key challenges in this task is the scarcity of annotated data, which limits the development and performance of effective retrieval models. In this technical report, we propose an audio-text retrieval model that aims to overcome the shortage of data and achieve competitive performance in audio-text multi-modal retrieval tasks.

To address the data scarcity issue, we propose two complementary approaches. The first approach involves gathering large paired audio-text datasets, where each audio sample is associated with corresponding textual information. By collecting a substantial amount of such paired data, we aim to enhance the model's ability to learn meaningful audio-text representations and improve its retrieval performance.

In addition to the collection of paired audio-text datasets, our second approach incorporates various augmentation techniques. Specifically, we employ techniques such as PairMix and multi-level test-time augmentation (Multi-TTA) to artificially expand the diversity and size of the training data. PairMix combines segments of audio from different pairs to create new instances, while Multi-TTA applies diverse transformations to the audio and text inputs during inference. These augmentation techniques enable the model to generalize better and improve its ability to handle unseen audio-text pairs during retrieval.

To evaluate the effectiveness of our proposed approaches, we conducted comprehensive experimental evaluations on the DCASE 2023 language-based audio retrieval challenge dataset. Our experiments demonstrate the benefits of both large paired audio-text datasets and augmentation techniques in enhancing the performance of the retrieval model. Moreover, our proposed model achieves competitive results in audio-text multi-modal retrieval tasks, highlighting its potential for practical applications in audio content retrieval.

The remainder of this technical report is organized as follows. Section 2 describes the methodology and architecture of our proposed audio-text retrieval model. Experimental setup and results are presented in Section 3, followed by a discussion of the findings in Section 4. Finally, Section 5 concludes the report and outlines future directions for research in this area.

2. SYSTEM DESCRIPTION

2.1. Augmentation

2.1.1. Uni-modal Augmentation

Our augmentation strategies for audio include applying Gaussian noise at the waveform level with a probability of 0.5. Spectrogram-level audio augmentation was also implemented using SpecAugment [1], which is incorporated in our baseline model. In addition, two strategies were used for text augmentation. We first applied Easy Data Augmentation (EDA) [2], which encompasses synonym replacement, random insertion, random swapping, and random deletion. Furthermore, we utilized ChatGPT[1] for paraphrasing captions, thereby enhancing the diversity of our captions. We used the method suggested in WavCaps [3] as a prompt to generate paraphrased captions. Note that the GPT-based caption replacement was only employed in AudioCaps and Clotho.

2.1.2. Multi-modal Augmentation: PairMix

In [4], PairMix, which mixes randomly selected raw images and concatenates their corresponding texts was introduced. It begins by randomly selecting N audio-text pairs $(a_i, t_i)_{i=1}^{N}$, each consisting of audio a_i and text t_i. It then mixes each modality separately to create a new audio-text pair. PairMix probabilistically applies either waveform-level or mel spectrogram-level audio mixup to produce a new mel spectrogram, denoted $\hat{s}$. Finally, PairMix constructs a new audio-text pair $(\hat{s}, \hat{t})$ by concatenating the text.

2.1.3. Test-time Augmentation: Multi-TTA

Test-time augmentation (TTA) can enhance the generalization of models by generating multiple predictions from augmented inputs and then averaging these predictions. Traditional TTA techniques only average outputs derived from augmented inputs. To exploit the full potential of TTA, we use a multi-level TTA (Multi-TTA) approach by generalizing the conventional TTA. Unlike traditional TTA which applies augmentations at a single layer, Multi-TTA selects multiple layers for TTA, thus allowing augmentations to be aggregated at different layers.

2.2. Dataset

Four datasets were used for training and evaluation.

Clotho Clotho v2 [5] is a dataset officially evaluated by challenge and used for training and evaluation. In both development and validation sets, five captions are assigned to one audio. The development set consists of 3,839 audio, 19,195 captions, and the validation set consists of 1,045 audio and 5,225 captions.

AudioCaps AudioCaps [6] is subset of Audioset [7] and is a dataset that has been annotated through crowd-sourcing of audio files. Train-set are used to train the model, and both audio and capture consist of 45,743 files each.

WavText5K The WavText5K [8] data was sourced from two websites: BigSoundBank[2] and SoundBible[3]. We used total of 3,685 audio and text parallel datasets for training model.

WavCaps WavCaps [3] is weakly-labelled audio captioning dataset created by ChatGPT. A total of 363,678 audio-text paired datasets were used to train the model.

2.3. Model

In alignment with previous research, our model employs an audio and language encoder architecture combined with a contrastive loss function. We utilize weights from the pre-trained audio neural networks (PANNs) [9] for the audio encoder. These networks have been trained on the AudioSet, which is currently the most extensive audio tagging dataset available. The PANNs, based on ResNet38, are capable of extracting broad representations from a wide range of audio clips. For the text encoder, we initialize our model with weights from the Bidirectional Encoder Representations from Transformers (BERT) [10]. The [CLS] token in the last layer serves as a text embedding. To further optimize our model, we allow both the audio and text encoders to be trainable by unfreezing them. Additionally, we incorporate two fully connected layers into the architecture to capture the relationship between the audio and text data.

3. EXPERIMENTS

3.1. Experimental Setup

Before training, all audio data were downsampled to a rate of 16000 Hz. During the pre-training phase which utilizes all four datasets, models were trained for 80 steps, with each step consisting of 540 data pairs. For fine-tuning which only uses Clotho dataset, the models were trained over 30 epochs, with a consistent batch size of 32 for all models. The AdamW optimizer, with a weight decay of 10^{-6}, was used and the learning rate was set to 10^{-5} for both the pre-training and fine-tuning phases. During pre-training, all models were exposed to Gaussian noise, SpecAugment, Easy Data Augmentation (EDA), ChatGPT-based caption paraphrasing, and PairMix. When it came to fine-tuning, we used two distinct approaches: one model applied only SpecAugment, while the other model employed all types of augmentations.

3.2. Result

We made four submissions to the challenge, which are detailed below:

- Submission 1: Models were pre-trained on all four datasets using uni-modal and multi-modal augmentations, then fine-tuned on Clotho using only SpecAugment. The test data underwent augmentation with SpecAugment, and the results were aggregated using multi-TTA. We performed middle-level aggregation on five samples and output-level aggregation on ten samples.
- Submission 2: Models were pre-trained on all four datasets using uni-modal and multi-modal augmentations and then fine-tuned on Clotho also with every augmentation applied in pre-training. The test data was aug-

[1] https://openai.com/blog/chatgpt

[2] https://bigsoundbank.com/

[3] https://soundbible.com/

mented with SpecAugment and the results were ensembled using multi-TTA, with middle-level aggregation at five samples and output-level aggregation at ten samples.

- Submission 3: Models were pre-trained on all four datasets using uni-modal and multi-modal augmentations, then fine-tuned on Clotho with only SpecAugment applied. For the test data, we applied SpecAugment and used multi-TTA to ensemble the results. Both middle-level and output-level aggregations were conducted at a single sample each.
- Submission 4: Models were pre-trained on all four datasets using uni-modal and multi-modal augmentations and then fine-tuned on Clotho also with every augmentation applied in pre-training. We augmented the test data with SpecAugment and used multi-TTA to ensemble the results, performing both middle-level and output-level aggregations at one sample each.

The outcomes for each submission are displayed in Table 1. All our models surpass the performance of the baseline model provided in the challenge. Upon comparing the models that used all types of augmentations in the Clotho fine-tuning (Submissions 2 and 4) with those that solely applied SpecAugment (Submissions 1 and 3), we found that the latter delivered superior results. A possible explanation for this is that the Clotho dataset is relatively small, and therefore the introduction of too many augmentations may inadvertently interfere with learning the inherent relationships present in the dataset.

4. CONCLUSIONS

Our technical report has introduced a novel audio-text retrieval model, specifically designed to address the DCASE 2023 language-based audio retrieval challenge. Recognizing the critical obstacle of data scarcity, we suggested two key approaches to solve this issue. The first strategy involved the collection of substantial paired audio-text datasets to create a more robust and representative database for our model to learn from. The second strategy applied various data augmentation techniques, notably PairMix and Multi-TTA, to increase the variability and overall richness of our dataset. Our empirical evaluations have demonstrated the validity and effectiveness of these strategies, attesting to their potential to substantially enhance the capabilities of audio-text retrieval models. Our model demonstrated robust performance across different audio-text multi-modal retrieval tasks.

5. ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions) and [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

Table 1: Evaluation of the model in development-testing split of Clotho.

	R@1	**R@5**	**R@10**	**mAP@10**
Challenge baseline	13.00	34.30	48.00	22.20
Submission 1	17.28	41.93	56.27	27.93
Submission 2	16.88	40.94	55.16	27.06
Submission 3	17.59	42.24	56.67	28.03
Submission 4	16.96	41.03	54.99	27.15

6. REFERENCES

[1] 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," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[2] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6382–6388.

[3] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[4] E. Kim, J. Kim, Y. Oh, K. Kim, M. Park, J. Sim, J. Lee, and K. Lee, "Improving audio-language learning with mixgen and multi-level test-time augmentation," *arXiv preprint arXiv:2210.17143*, 2022.

[5] 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.

[6] 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.

[7] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[8] S. Deshmukh, B. Elizalde, and H. Wang, “Audio retrieval with wavtext5k and clap training,” *arXiv preprint arXiv:2209.14275*, 2022.

[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. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pretraining of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.