TEXT-TO-AUDIO RETRIEVAL: ENSEMBLE COMBINATIONS OF THE MODELS.

Technical Report

Jiwon Park[1], *SangJe Park*[1], *Changwon Lim*[1]

[1]Chung-Ang University, Department of Applied Statistics, Seoul, South Korea,
{jiwon3401, pks5034, clim}@cau.ac.kr

ABSTRACT

This technical report focuses on the audio-text retrieval model designed for the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2023 Task 6b. In this task, the objective is to retrieve 10 audio files from a given dataset based on a given text query and then sort them according to how well they match the query. The audio encoder in our model employs Pretrained Audio Natural Networks (PANNS), which is a pre-trained model from the AudioSet dataset. We have fine-tuned the encoders using the Clotho dataset. For the text encoder, we have used transfer learning with Sentence-BERT, which is based on the Transformer architecture. To bring audio and text inputs into a joint embedding space, we have passed them through their respective encoders. We have then employed contrastive learning for audio-text pairs so that similar pairs are positioned close together and the other pairs are positioned further apart. We achieves 0.245 on mAP10 of text-to-audio retrieval.

***Index Terms*—** Audio Retrieval, Text to Audio Retrieval, Contrastive learning

1. INTRODUCTION

Audio retrieval is a cross-modal retrieval task to identify the corresponding audio or text when one of them is given as input. It involves generating new representations from different modalities and mapping them to a shared subspace. The purpose of text-to-audio retrieval is to get 10 audio files from a specified dataset for each text query and sort them depending on how well they fit the query. For instance, given a text query such as ”Loud pops of rain splashing down on the ground ”, the goal of the task is to calculate relevance scores of audio samples with respect to a given caption query and subsequently sort the audio samples in descending order based on their relevance scores. [1] investigates the importance of various metric learning objectives on the audio-text retrieval task. We conducted experiments using various combinations of contrastive learning losses and encoders to evaluate their effectiveness in the audio-text retrieval task.

2. PROPOSED METHOD

2.1. System Overview

Figure 1 shows the proposed system overview.

The audio and text features are extracted using an audio encoder and a text encoder, which are then projected onto a unified embedding space. In order to align audio and text inputs within a unified embedding space, we utilized separate encoders for audio modality. By passing the audio and text inputs through their respective encoders, we obtained corresponding representations. To train the model, we employed contrastive learning techniques on the audio-text pairs.

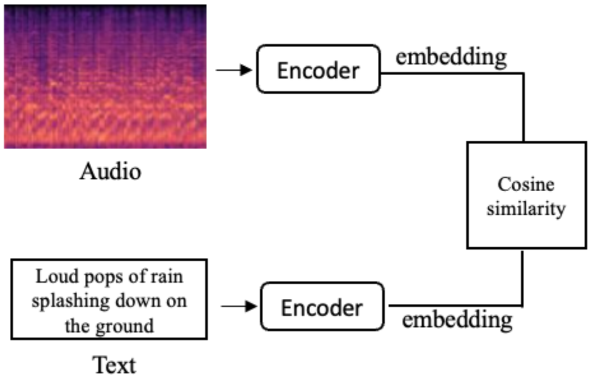

Figure 1: Model Architecture

2.2. Audio Encoder

Our model utilizes the Audio Encoder from pre-trained audio neural networks (Panns) [2], which comprises three different encoders: CNN14, Resnet38, and Wavegram-logmel-Cnn14. We used the encoders for the feature extracting of input log-mel spectrogram. Wavegram-logmel-Cnn14 encoder leverages data from both log mel spectrograms and time-domain waveforms. This is achieved through a combination along the channel dimension, allowing the model to capture complementary information from both representations.

2.3. Text Encoder

The Text Encoder utilized in our model is Sentence-BERT (SBERT) [3], which is an enhanced version of the BERT network specifically designed for generating high-quality sentence embeddings. S-BERT incorporates siamese and triplet network architectures to derive semantically meaningful representations of sentences. These embeddings can be compared using cosine similarity, enabling efficient semantic similarity calculations between sentences. Our system benefits from improved performance in generating robust and context-aware sentence embeddings for text-based tasks.

3. EXPERIMENTS

3.1. Dataset

Clotho v2.1 dataset [4] consists of 3839, 1045, and 1045 audio clips in the training, validation, and test sets respectively, while the development set comprises 5929 audio clips, each associated with five reference captions.

3.2. Experiment setups

In the training process, epoch 20 and batch size of 96 is used with a learning rate of 10^{-4}. We use the Adam Optimizer [5] and ReduceLROnPlateau learning rate schedular. The experiments were conducted at various random seeds. The audio feature extracted from log-mel spectrogram is obtained by sampling rate 32kHz, Hanning window of 1024 with 64 mel bins. Spec Augment [6] is employed as a data augmentation method, applying frequency and time masks to the log-mel spectrogram input to enhance training robustness.

For comparative experimentation, the model is trained using six different loss functions: Triplet-sum, Triplet-max [7], Triplet-weighted [8], NT-Xent [9], InfoNCE [10], VICReg [11].

InfoNCE is widely used in contrastive learning and VICReg is used for preventing mode collapse problem. [12] used them together so that they can improve performance while preventing collapse.
4 models of the highest mAP10 score are selected for submission:

- Ensemble of 5 models: different architectures with different loss
- Ensemble of 4 models : different architectures with different loss
- Ensemble of 3 models: same architecture with 3 different triplet loss
- Ensemble of 2 models: same architecture with InfoNCE loss and InfoNCE+VICReg

4. RESULTS

The performance of audio retrieval is shown in Table 1.

Model	loss	R@1	R@5	R@10	mAP10
Cnn14 + SBERT	Triplet-weighted	13.86	36.08	49.24	23.54
	Triplet-max	13.74	35.25	48.44	23.00
	NT-Xent	13.40	35.29	48.61	22.87
	InfoNCE + VICReg	10.58	30.62	43.29	19.30
Resnet38 + SBERT	Triplet-weighted	13.21	34.11	47.14	22.33
	NT-Xent	12.38	33.88	46.49	21.56
WLCNN14 + SBERT	Triplet-weighted	11.87	33.42	46.35	21.31
Ensemble		14.74	37.59	50.68	24.46

Table 1: Score for model performance on evaluation data

5. CONCLUSION

We experimented that using pre-trained PANNs and Sentence-Bert for audio and text encoders each can show consistent performance when trained with the most layers except the projector frozen. We conducted various ensemble combinations of audio encoder (Cnn14, ResNet38, WLCNN14) and loss (Triplet-weighted, Triplet-max, NT-Xent, InfoNCE, VICReg) and found out that the ensemble of consistent models benefit robustness.

6. REFERENCES

[1] 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.

[2] 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*, pp. 2880–2894, 2020.

[3] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 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] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[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] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” *arXiv preprint arXiv:1707.05612*, 2017.

[8] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, “Universal weighting metric learning for cross-modal matching,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 005–13 014.

[9] 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.

[10] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018. [Online]. Available: https://arxiv.org/abs/1807.03748

[11] A. Bardes, J. Ponce, and Y. LeCun, “Vicreg: Variance-invariance-covariance regularization for self-supervised learning,” *arXiv preprint arXiv:2105.04906*, 2021.

[12] T. Lepage and R. Dehak, “Label-efficient self-supervised speaker verification with information maximization and contrastive learning,” *arXiv preprint arXiv:2207.05506*, 2022.