A RESNET-BASED CLIP TEXT-TO-AUDIO RETRIEVAL SYSTEM FOR DCASE CHALLENGE 2022 TASK 6B

Technical Report

Yongquan Lai, Jinsong Pan, Buxian Chen

PingAn Property & Casualty Insurance Company of China, Ltd.

ABSTRACT

Language-based audio retrieval aim to use language to retrieval audios in a given dataset. This technical report presents a text-to-audio retrieval system submitted to Task 6b of the DCASE 2022 challenge. The proposed system is based on AudioCLIP, which incorporates the ESResNeXt audio-model into the CLIP framework using the AudioSet and clothe V2 datasets and introduces a pre-training method to perform bimodal querying. the original AudioCLIP acquired poor retrieval performance on the clothe V2 dataset in a zero-shot inference fashion. So we used AudioCLIP's model as a weight initializer, and finetuned audio encoder and text encoder using symmetric cross entropy loss over similarity measure among the mini-batch (audio, text) pairs. Through pre-training and data augmentation methods, our model achieved R1 score of 0.35 and mAP10 score of 0.51 on Clotho V2 evaluation set.

Index Terms— audio retrieval, cross-modal task, pre-training, data augmentation

1. INTRODUCTION

Given a caption as a query, text-audio retrieval aims to retrieving best matched audios from a pool of candidates. This cross-modal retrieval task is challenging as it requires not only learning robust feature representations for both acoustic and textual modalities but also capturing fine-grained interaction between the learned acoustic and textual features and aligning them in a shared embedding space. This task has received extensive attention in recent years [2, 3, 6, 7].

The AudioCLIP architecture was shown to give excellent querying performance on UrbanSound8K and ESC-50 [2], and thus is chosen as the baseline system in our work. Language-based audio retrieval requires to extract features from the audio and text into a shared feature space and get the top-k matching scores. The Clotho V2 dataset provided by the website is limited, so we used the pre-training method to improve model performance. The query model was pre-trained on the AudioSet dataset and fine-tuned on the clothe V2 development dataset. All results were compared on the clothe V2 evaluation dataset.

2. SYSTEM DESCRIPTION

The proposed system is based on AudioCLIP, which incorporates the ESResNeXt audio-model into the CLIP framework. As is shown in Figure 1, the ESResNeXt model is responsiable for the audio encoding part; the CLIP model is responsiable for the caption encoding part.

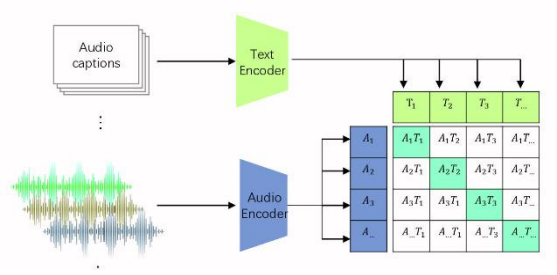

Figure 1. Overview of the proposed AudioCLIP model.

2.1 Audio Encoder

For the audio encoding part, The ESResNet model [8] combined commonly used visual domain techniques such as a ResNet-based backbone, Siamese-like multi-channel processing, and depth-wise separable convolutions together with the computation of log-power spectrograms obtained using Short-Time Fourier Transform and cross-domain transfer learning in application to downstream tasks. An overview of the ESResNet's processing pipeline is given by Figure 2. The chosen model contains moderate number of parameters to learn (~30M). the ESResNeXt model supports an implicit processing of a multi-channel audio input and provides improved robustness against additive white Gaussian noise and sample rate decrease.

2.2 Text Encoder

The original CLIP model consists of two subnetworks: text and image encoding heads. the proposed model only contains text encoder. The text encoder is a Transformer with the architecture modifications [9]. As a base size we use a 12-layer 512-wide model with 8 attention heads. The transformer operates on a lower-cased byte pair encoding (BPE) representation of the text. The text sequence is bracketed with [SOS] and [EOS] tokens and the activations of the highest layer of the transformer at the [EOS] token is used as the feature representation of the text which is layer normalized and then linearly projected into the multi-modal embedding space. Masked self-attention was used in the text encoder to preserve the ability to add language modeling as an auxiliary objective [7].

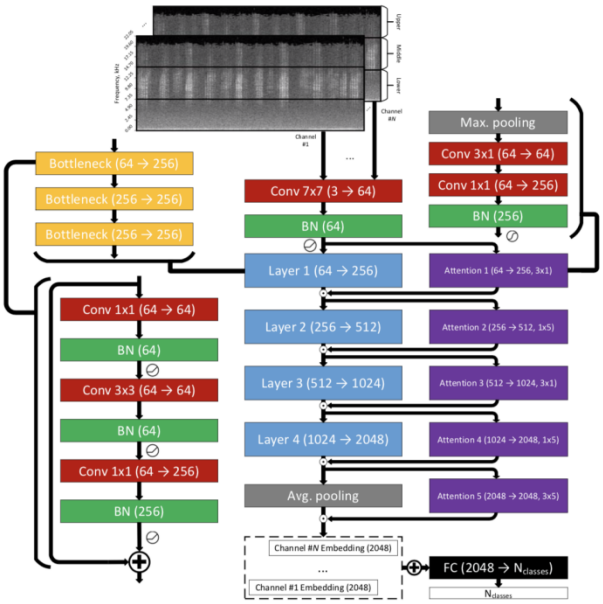

Figure 2. Overview of the ESResNet model.

2.3 Pretrained model for encoder

The pre-trained model was trained by AudioCLIP. While the CLIP model was already pre-trained on text-image pairs, the pre-training model perform an extended AudioSet pre-training of the audio-head first, as it improved performance of the base ESResNeXt model, and then to continue training in a tri-modal setting combining it with two other heads. Here, the whole AudioCLIP model was trained jointly on the AudioSet dataset using audio snippets, the corresponding video frames and the assigned textual labels.[2]

3. EXPERIMENTS

3.1 Dataset

Clotho V2 is an audio captioning dataset containing a total of 6974 audio samples collected from the Freesound platform and annotated on Amazon Mechanical Turk by annotators from English-speaking countries. To encourage caption diversity, each audio clip is provided with 5 captions annotated by different annotators, thus there are in total 34870 captions. The duration of the audio samples is uniformly ranged from 15 to 30 seconds. Captions are post-processed to make sure there are no unique words, named entities and speech transcriptions.

The Clotho V2 evaluation set is reserved as the evaluation set for the DCASE challenge. Audio clips in the development set are randomly sampled to create a training set with 5719 audio samples and a validation set with 200 audio samples. During training, each audio clip is combined with one of its five captions as a training sample. During evaluation, all five ground truth captions of an audio clip are used as references and compared with the predicted caption for metric computation.

3.2 Data pre-processing

The finetuned model was also trained by AudioCLIP on Clotho V2 dataset. The captions input directly to text encoder. The audios were clipped to 20s and then input to encoder.

3.3 Experimental setups

The whole model is trained using the Adam optimizer with a batch size of 6 on V100. The learning rate linearly increased to 1e-5. Dropout with rate of 0.2 is applied in the proposed model to mitigate over-fitting problems. To improve the generalization ability of the model, label smoothing is applied in all the experiments. used AudioCLIP's full training model as a weight initializer, and finetuned audio encoder and text encoder using symmetric cross entropy loss over similarity measure among the mini-batch (audio, text) pairs.

The model is trained for 40 epochs, and the top-5 models are taken as the final model after the weight average operation.

3.4 Results

In this work, we compared the retrieval ability of different models on Clotho V2 evaluation set. The baseline results are from DCASE official website [10]. The details of the retrieval models in the Table 1 are as follows:

- AudioCLIP: the origin full training model, which achieves state-of-the-art results in the Environmental Sound Classification task.
- AudioCLIP-F1: This model is first pre-trained on AudioCaps and then fine-tuned on Clotho V2. During finetuning step, the audios randomly intercepted for 10s in the time domain and are integrated into a full matrix, then are entered into the audio encoder.
- AudioCLIP-F2: This model is first pre-trained on AudioCaps and then fine-tuned on Clotho V2. During finetuning step, the audios randomly intercepted for 20s in the time domain and are integrated into a full matrix, then are entered into the audio encoder. Note that when the duration of audio is shorter than 20s, the end of the audio feature needs to be padded.

Table 1. The results of models on clothe V2 evaluation dataset

model	R1	R5	R10	mAP10
baseline	0.03	0.11	0.19	0.07
AudioCLIP	0.04	0.12	0.17	0.42
AudioCLIP-F1	0.20	0.59	0.64	0.37
AudioCLIP_F2	0.35	0.89	1.00	0.51

The challenge allows us to submit up to four different results. Our submission contains four results, which are come from AudioCLIP-F2 model.

4. CONCLUSION

This technical report briefly describes our system and methods for Task 6B of DCASE 2021. Using pre-training learning, the proposed system significantly improves all evaluation metrics compared to the top-ranked systems in the DCASE challenge last year.

5. REFERENCES

[1] http://dcase.community/workshop2022/.

[2] A. Guzhov, F. Raue, J. Hees and A. Dengel, AudioCLIP: Extending CLIP to Image, Text and Audio. arXiv preprint-arXiv:2106.13043, 2021.

[3] Guzhov A, Raue F, Hees J, et al. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[4] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[5] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[6] Gabeur V, Sun C, Alahari K, et al. Multi-modal transformer for video retrieval[C]//European Conference on Computer Vision. Springer, Cham, 2020: 214-229.

[7] Mei X, Liu X, Sun J, et al. On Metric Learning for Audio-Text Cross-Modal Retrieval[J]. arXiv preprint arXiv:2203.15537, 2022.

[8] Guzhov A, Raue F, Hees J, et al. Esresnet: Environmental sound classification based on visual domain models[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 4933-4940.

[9] Solaiman I, Brundage M, Clark J, et al. Release strategies and the social impacts of language models[J]. arXiv preprint arXiv:1908.09203, 2019.

[10] https://dcase.community/challenge2022/task-language-based-audio-retrieval