DCASE 2022 CHALLENGE TASK 6B: LANGUAGE-BASED AUDIO RETRIEVAL

Technical Report

Huang Xie, Samuel Lipping, Tuomas Virtanen

Audio Research Group, Tampere University, Tampere, Finland

ABSTRACT

In this report, we introduce the task setup and the baseline system for the sub-task B of the DCASE 2022 Challenge Task 6: language-based audio retrieval subtask. For this subtask, the Clotho v2 dataset is utilized as the development dataset, and an additional dataset consisting of 1,000 audio-caption pairs as the evaluation dataset. We train the baseline system with the development dataset, and evaluate it on the evaluation dataset to provide some initial results for this subtask.

Index Terms— Language-based audio retrieval, cross-modal learning, audio captioning, Clotho.

1. INTRODUCTION

With the growth of multimedia content in recent decades, there is a need for retrieval methods that can efficiently organize the data based on its content, and retrieve relevant items when doing searches to datasets. Natural language provides an efficient way to represent complex information about multimedia. It can represent high level information about data that goes beyond any fixed taxonomies. For audio signals, natural language can represent information related to temporal and spatial relationships between sound sources, and attributes of sounds and their environment.

Language-based multimedia retrieval has received increasing attention in recent years. The majority of recent works has focused heavily on the visual domain [1][2]. For example, there are plenty of proposed approaches [3] tackling content-based image retrieval with free-form textual descriptions. In contrast, only a few studies have been conducted on language-based audio retrieval in the existing literature. Early works [4][5] deal with language-based audio retrieval using multi-word text queries consisting of audio tags or class labels, rather than sentence-like textual descriptions, e.g., captions. Recent studies [6][7] prompt research in this field by exploring human written caption as queries. Through the sub-task B of the DCASE 2022 Challenge Task 6, we aim to inspire further research into language-based audio retrieval with unconstrained textual descriptions.

2. TASK DESCRIPTION

The goal of language-based audio retrieval is to evaluate audio retrieval methods, where a retrieval system takes a free-form textual description as an input, and the retrieval system is supported to rank audio signals in a fixed dataset based on their match to the given description. In DCASE 2022 Challenge, human written audio captions will be used as input queries. For each query, the retrieval task is to retrieve 10 audio files from a given evaluation dataset and sort them according to their match with the query.

### 2.1. Development Dataset

The development dataset for this task is Clotho v2 [8], which consists of audio samples of 15 to 30 seconds duration, with each audio sample having five captions of eight to 20 words length. There are 6,974 audio samples with 34,870 captions in total. All audio samples are sourced from the Freesound platform [9], and captions are crowd-sourced using a three-step framework [8].

The Clotho v2 dataset is divided into a training split of 3,839 audio clips with 19,195 captions, a validation split of 1,045 audio clips with 5,225 captions, and a testing split of 1,045 audio clips with 5,225 captions. These splits are created by first constructing the sets of unique words of the captions of each audio clip. These sets of words are combined to form the bag of words of the whole dataset, from which the frequency of a given word can be derived. With the unique words of audio files as classes, multi-label stratification is applied. The data collecting procedure is explained in detail in [8].

#### 2.2. Clotho Retrieval Evaluation Dataset

The evaluation dataset for this task consists of 1,000 audio samples sourced from the Freesound platform [9], and one human written caption is provided for each audio sample. The audio samples are collected following the procedure described in [8], by optimizing the tag distribution of the selected samples. The samples were selected from the set of files not used in Clotho v2. Captions were gathered using the first crowdsourcing step of Clotho v2 manually screened for typographical errors and speech transcription. Table 1 summarizes the information about the development and evaluation datasets.

| Dataset   | Split | #Audio | #Captions |
|-----------|-------|--------|-----------|
| Development | Training | 3839   | 19195     |
|           | Validation | 1045   | 5225      |
|           | Testing   | 1045   | 5225      |
| Evaluation |          | 1000   | 1000      |

Table 1: Statistics of the development and evaluation datasets.

#### 2.3. Evaluation Metrics

In the evaluation, the ground truth relevancy of audio samples are considered binary (i.e., only the audio samples belong to the caption query are considered relevant, and all the others not relevant). The submissions for this task will be evaluated using mean average precision at top-10 (mAP@10) as the main metric, and recall at $k$ (R@k with $k \in \{1, 5, 10\}$) as the secondary metrics.
The mAP metric has been widely used for evaluating the performance of cross-modal retrieval algorithms [3]. It is a rank-aware metric, which measures the mean of average precisions (AP) over all the queries. The AP for a query is calculated by averaging the precisions at positions, where relevant items are in the retrieved rank list. The more relevant items in the top rank list, the higher mAP value it has. The R@k metric is another standard, rank-unaware retrieval metric [6], which is defined as the proportion of relevant items among the top-K retrieved results to all the relevant items in the evaluation dataset, averaged across all the caption queries. The submissions for this task will be ranked by the mAP@10 metric.

3. BASELINE SYSTEM

To provide a starting point and some initial results for the challenge, we present a baseline system for the language-based audio retrieval subtask.

3.1. Baseline System

The baseline system is a simplified version of the audio-text aligning framework presented in [7], which calculates relevant scores between encoded textual descriptions (i.e., encoded captions) and encoded audio signals. It consists of two input encoders: one for audio, and the other for text, as illustrated in Figure 1. These two modality-specific encoders generate vector representations (i.e., audio embeddings and caption embeddings) for audio clips and textual descriptions. Then, the relevant score between an audio clip and a textual description is calculated by the dot product of their vector representations. The baseline system is optimized with a sampling-based triplet loss [10] at the training stage, and then applied to retrieve audio for caption queries at the testing stage.

3.2. Audio Encoder

A convolutional recurrent neural network (CRNN) [11] is used as the audio encoder, which extracts frame-wise acoustic embeddings from audio signals. It consists of five convolution blocks, followed by a bidirectional gated recurrent unit (BiGRU). Each convolution block includes an initial batch normalization, a convolutional layer with padded 3 × 3 convolutions, and a LeakyReLU activation with a slope of −0.1. After the first, third, and fifth convolution blocks, one L4-Norm subsampling layer is used to reduce the temporal dimension of each block’s output by a factor of four. A dropout layer with a rate of 0.3 is placed between the last L4-Norm layer and the BiGRU. Lastly, an up-sampling operation is applied to ensure the final output has the same temporal dimension as the CRNN input.

The CRNN audio encoder takes 64-dimensional log mel-band energies as input. Each audio clip is split into 40 ms Hanning-windowed frames with a hop length of 20 ms. Then, 64 log mel-band coefficients are extracted from each frame. A sequence of 300-dimensional frame-wise acoustic embeddings are generated for each audio clip. The final audio embedding is calculated by averaging the frame-wise acoustic embeddings.

3.3. Text Encoder

Word2Vec (Skip-gram model) [12] is utilized as the text encoder to convert textual descriptions into sequences of word embeddings. It is a two-layer fully-connected neural network, which learns word embeddings that are good at predicting surrounding words in a sentence or a document. For the sake of simplicity, we adopt a publicly available pre-trained Word2Vec [13], which is trained on Google News dataset. It consists of 300-dimensional word embeddings for roughly three million case-sensitive English words and phrases. The Word2Vec text encoder converts textual descriptions into sequences of semantic word embeddings word by word. The final caption embedding is computed by averaging the word embeddings.

3.4. Training Objective

We train the baseline system by optimizing a ranking-based criterion [10], such that audio clips and captions that belong together are more similar in the embedding space than mismatched audio-caption pairs. Specifically, across a batch of N audio-caption pairs \( \{(x_n, y_n)\}_{n=1}^{N} \), where \( y_n \) is the caption pertaining to an audio clip \( x_n \), we randomly select an imposter clip \( \hat{x}_n \) and an imposter caption \( \hat{y}_n \) for each audio-caption pair \( (x_n, y_n) \). Then, we calculate the widely used sampling-based triplet loss [7,14,15]

\[
loss = \frac{1}{N} \sum_{n=1}^{N} \left[ \max(0, S(x_n, \hat{y}_n) - S(x_n, y_n) + \eta) + \max(0, S(\hat{x}_n, y_n) - S(x_n, y_n) + \eta) \right],
\]

where \( S \) is the audio-caption relevant score and \( \eta \) is a margin hyper-parameter. We fix \( \eta \) to one.
Table 2: Experimental results on the testing split and the evaluation dataset.

| Dataset       | mAP@10 | R@1  | R@5  | R@10 |
|---------------|--------|------|------|------|
| Testing split | 0.07   | 0.03 | 0.11 | 0.19 |
| Evaluation    | 0.06   | 0.03 | 0.10 | 0.18 |

4. EXPERIMENTAL RESULTS AND DISCUSSION

We train the baseline system with the training and validation splits of the development dataset, and report the experimental results on the testing split and the evaluation dataset.

4.1. Experimental Setup

We train the baseline system with batches of 32 audio-caption pairs in the training split for at most 150 epochs, and monitor the loss on the validation split during the training process. An Adam optimizer with an initial learning rate of 0.001 is adopted to optimize the training process. The learning rate is reduced by a factor of ten once the validation loss does not improve for five epochs. Training is terminated by early stopping with ten epochs.

4.2. Results

As shown in Table 2, the baseline system achieves similar performance in terms of mAP@10 and recall scores on the testing split and the evaluation dataset. Specifically, with the evaluation dataset, the theoretical chance levels are 1/1000 = 0.001 for R@1, 1/200 = 0.005 for R@5, and 1/100 = 0.01 for R@10, respectively. In contrast to the theoretical chance levels, the baseline system obtains better recall scores, with R@1 around 0.03, R@5 around 0.10, and R@10 around 0.18. The experimental results show that the baseline system can retrieve audio with their corresponding captions, i.e., perform language-based audio retrieval. On the other hand, since the baseline system employs a simple pipeline (e.g., averaging frame-wise acoustic embeddings and word embeddings), the retrieval performance could be limited.

5. CONCLUSIONS

In conclusion, we introduce the task setup, including the task description, the development and evaluation datasets, and evaluation metrics, for the DCASE 2022 Challenge Task 6b: language-based audio retrieval subtask. We also present a baseline system to provide some initial results for this subtask. The experimental results on the testing split of the development dataset and the evaluation dataset show that the baseline system can retrieve audio with textual descriptions.

6. REFERENCES

[1] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang, “Dual encoding for video retrieval by text,” IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–10, 2021.

[2] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, “Dynamic modality interaction modeling for image-text retrieval,” in Proc. ACM SIGIR Int. Conf. Res. Dev. Inf. Retr., 2021, pp. 1104–1113.

[3] P. Kaur, H. S. Pannu, and A. K. Malhi, “Comparative analysis on cross-modal information retrieval: A review,” Comput. Sci. Rev., p. 100336, 2021.

[4] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon, “Largescale content-based audio retrieval from text queries,” in Proc. Int. Conf. Multimed. Inf. Retr., 2008, pp. 105–112.

[5] B. Elizalde, S. Zarar, and B. Raj, “Cross modal audio search and retrieval with joint embeddings based on text and audio,” in Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP), 2019, pp. 4095–4099.

[6] A. M. Oncescu, A. Sophia Koepke, J. F. Henriques, Z. Akata, and S. Albanie, “Audio retrieval with natural language queries,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2021, pp. 2411–2415.

[7] H. Xie, O. Räisänen, K. Drossos, and T. Virtanen, “Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases,” 2022, accepted at ICASSP.

[8] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: an audio captioning dataset,” in Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP), 2020, pp. 736–740.

[9] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in Proc. Int. Conf. Multimed., 2013, pp. 411–412.

[10] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a ”siamese” time delay neural network,” in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), 1993, pp. 737–744.

[11] X. Xu, H. Dinkel, M. Wu, and K. Yu, “A crnn-gru based reinforcement learning approach to audio captioning,” in Proc. Detect. Classif. Acoust. Scenes Events Work. (DCASE), 2019, pp. 225–229.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. Learn. Representations (ICLR), 2013.

[13] “Word2Vec,” https://code.google.com/archive/p/word2vec, accessed: 2021-09-27.

[14] D. Harwath, A. Torralba, and J. R. Glass, “Unsupervised learning of spoken language with visual context,” in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), 2016, pp. 1866–1874.

[15] D. Harwath, W. Hsu, and J. R. Glass, “Learning hierarchically discrete linguistic units from visually-grounded speech,” in Proc. Int. Conf. Learn. Representations (ICLR), 2020, pp. 1–22.