LEVERAGING MULTI-TASK TRAINING AND IMAGE RETRIEVAL WITH CLAP FOR AUDIO CAPTIONING

Technical Report

Haoran Sun, Zhiyong Yan, Yongqing Wang, Heinrich Dinkel, Junbo Zhang, Yujun Wang

Xiaomi Corporation, Beijing, China
{sunhaoran7,yanzhiyong,wangyongqing3,dinkelheinrich,zhangjunbo1,wangyujun}@xiaomi.com

ABSTRACT

This technical report serves as our submission to Task 6 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 challenge. Our system, as described in this report, consists of two sub-systems designed for the respective sub-tasks: automated audio captioning (task A) and text-to-audio retrieval (task B).

The text-to-audio retrieval system employs a tri-encoder architecture, where pre-trained audio and text encoders are trained to establish relationships. Additionally, an extra pre-trained image encoder is utilized to enhance the connections between these encoders. Through this retrieval process, the audio encoder can be considered a pre-trained encoder for task A.

Furthermore, we employ multi-task training with audio tagging during the retrieval phase to strengthen the encoder for audio captioning. Pre-training is conducted using AudioCaps and a portion of WavCaps datasets, and both sub-systems are subsequently fine-tuned on Clotho dataset.

Experimental results demonstrate that our model achieves a SPIDEr score of 0.305 and a SPIDEr-FL score of 0.294 for captioning, as well as an mAP (mean Average Precision) of 0.321 for text-to-audio retrieval.

Index Terms— audio captioning, text-to-audio retrieval, pre-training, multi-task learning

1. INTRODUCTION

Audio captioning refers to the task of generating textual descriptions or captions for audio content. Similar to image captioning, which generates descriptions for images, audio captioning aims to provide textual representations that capture the key information, context, and semantic meaning of audio signals. Notable initial works in the field include [1, 2, 3], which proposed datasets and baselines that led to the creation of Task 6 within DCASE.

In recent years, audio captioning has gained increasing attention due to its potential applications in various domains, such as enhancing accessibility for individuals with hearing impairments by providing textual representations of audio content. More recently, audio-caption and its respective datasets have been the foundation of contrastive language audio pretraining (CLAP), with notable works being [4, 5, 6]. Furthermore, audio captioning serves as a valuable tool for audio-based media retrieval, enabling users to search and retrieve specific audio content using text-based queries or vice versa.

Audio caption is commonly modelled as a sequence-to-sequence problem: Given an audio signal, the goal is to generate a sentence that accurately describes the audio contents. Most successful approaches for audio captioning are based on an encoder-decoder architecture. An audio encoder is first fed the audio signal, which predicts rich, high-level embeddings, which are then further fed into a decoder that produces text. There has been a multitude of audio encoders [7, 8] and text-decoders [9, 10] investigated. Further, novel evaluation metrics that accurately reflect an audio caption model's performance have also been proposed [11, 12].

This paper is structured as follows. In Section 2 we introduce our methodology. Then in Section 3 we display our experimental setup and show the corresponding results in Section 4. Finally, Section 5 concludes the work.

2. METHOD

2.1. Text-to-Audio Retrieval

2.1.1. Contrastive Language–Audio Pre-training (CLAP)

The bi-encoder architecture is adapted from last year's winner [13], which consisting of an audio encoder E_A, a text encoder E_T and a cross-modal matching module. The encoders can transform an audio-text pair $(\mathcal{A}, \mathcal{T})$ into an embedding pair (e_a, e_t), then the cross-modal matching modules $Match_A$ and $Match_B$ will project them into a common space:

$$\begin{aligned} e_a &= E_A(\mathcal{A}), \\ e_t &= E_T(\mathcal{T}), \\ a &= Match_A(e_a), \\ t &= Match_T(e_t). \end{aligned} \tag{1}$$

The similarity score (cosine similarity in this system) between a and t can be obtained by:

$$s_{\mathcal{A}\sim\mathcal{T}} = \frac{a_p \cdot t_p^T}{\|a_p\| \cdot \|t_p\|} \tag{2}$$

The InfoNCE loss [14] is adopted as the training loss. This contrastive training loss between the similarity scores and the ground truth labels can be calculated as below:

$$\begin{aligned} \mathcal{L}_i^{A\longrightarrow T} &= -\log \frac{\exp(s_{\mathcal{A}\sim\mathcal{T}}(i,i)/\tau)}{\sum_{j=1}^{N}\exp(s_{\mathcal{A}\sim\mathcal{T}}(i,j)/\tau)}, \\ \mathcal{L}_i^{T\longrightarrow A} &= -\log \frac{\exp(s_{\mathcal{A}\sim\mathcal{T}}(i,i)/\tau)}{\sum_{j=1}^{N}\exp(s_{\mathcal{A}\sim\mathcal{T}}(j,i)/\tau)}, \\ \mathcal{L}_{CLAP} &= \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i^{A\longrightarrow T} + \mathcal{L}_i^{T\longrightarrow A}), \end{aligned} \tag{3}$$

where τ is the trainable temperature.

In our work, the model architectures are the same as Xu's [13], including CNN14 in PANNs, BERT [15] for E_T.

2.1.2. Contrastive Language–Audio-Image Pre-training (CLAIP)

In this paper, we utilized the stable diffusion model to generate image files corresponding to the training text. Building upon the CLAP training framework, we incorporated a contrastive loss between text embeddings and image embeddings to complement the learning of contrastive loss between text and audio. For image embedding extraction, we primarily employed the Vision Transformer[16, 17, 18] model from the LAION open-source repository[6].

The similarity score between text and image can be obtained the same as Formula 2, then the whole loss function for CLAIP can be calculate by:

$$\begin{aligned} \mathcal{L}_i^{I \longrightarrow T} &= -\log \frac{\exp(s_{\mathcal{I} \sim \mathcal{T}}(i,i)/\tau)}{\sum_{j=1}^{N} \exp(s_{\mathcal{I} \sim \mathcal{T}}(i,j)/\tau)}, \\ \mathcal{L}_{CLAIP} &= \mathcal{L}_{CLAP} + \frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_i^{I \longrightarrow T}). \end{aligned} \tag{4}$$

2.2. Audio Captioning

2.2.1. Audio Captioning with CLAP or CLAIP

The audio encoder obtained in Section 2.1 can be regarded as a feature extractor in audio captioning, which is denoted by E_A. Then, an audio encoder E_{AC}, a decoder D_{AC} and a fully connected classifier form the trainable layers to output the word probability.

$$\begin{aligned} e_a &= E_A(\mathcal{A}) \\ e_{ac} &= E_{AC}(e_a) \\ y &= D(e_{ac}, WE(w)) \\ o &= Classifier(y) \end{aligned} \tag{5}$$

where the word embedding layer WE extracts the embedding by the paired word w.

Cross entropy loss between the estimated word probability p and ground truth word w_t is adopted to optimize the entire model except for E_A, which is frozen while training.

$$\mathcal{L}_{AC} = -\frac{1}{T} \sum_{t=1}^{T} logp(w_t) \tag{6}$$

Pre-trained E_As in 2.1 are adopted as feature extractors, a three-layer bidirectional gated recurrent unit (GRU) is taken as E_{AC}, and a two-layer Transformer is taken as D_{AC}. Besides, we also tried to replace the E_a by the optimized one for this task described in Section 2.2.2.

2.2.2. CLAP with Audio Tagging for Audio Captioning

For the audio captioning feature extractor, we adopted audio encoders after performing audio-to-text retrieval. However, the retrieval training process focuses on establishing a relationship between audio embeddings and text embeddings, while overlooking the classification ability of the audio tagging model. Therefore, we propose incorporating multi-task training with audio tagging alongside the retrieval training. This approach allows us to maintain stronger classification capabilities while improving the embedding relationships.

The pre-trained audio-encoder has the most classification capabilities in the retrieval training process, so, we just need to reduce the difference between the output scores of the pre-trained encoder ($scores_{pre}$) and the training encoder ($scores_{tr}$):

$$\begin{aligned} \mathcal{L}_{TAG} &= loss_{reg}(scores_{tr}, scores_{pre}), \\ \mathcal{L}_{CLAP-TAG} &= \mathcal{L}_{CLAP} + \lambda \mathcal{L}_{TAG}. \end{aligned} \tag{7}$$

In the equation, $loss_{reg}$ represents the regularization loss function, which can be either the *Mean Absolute Error* (MAE) or the *Mean Squared Logarithmic Error Loss* (MSLE). The parameter λ denotes the weight assigned to the regularization loss.

Using the MSLE loss is advantageous because it operates on the same logarithmic scale as the InfoNCE loss used in retrieval processing. This compatibility between the losses leads to improved training performance and better overall results.

Furthermore, we explored the integration of the multi-task training method with CLAIP, with the objective of further enhancing the overall performance. The rationale behind this approach is to leverage the benefits of both techniques synergistically:

$$\mathcal{L}_{CLAIP-TAG} = \mathcal{L}_{CLAIP} + \lambda \mathcal{L}_{TAG} \tag{8}$$

3. EXPERIMENTS

3.1. Data

Clotho v2.1[19] is used as the dataset for both sub-tasks. We re-split the original training and validation sets into new subsets in a $9:1$ ratio. Then, there are 4395, 489, 1045 audio clips in the training, validation and evaluation sets. The re-splitting is used to get more data for training. Whole of pre-trained models mentioned below will be fine-tuned on Clotho.

For both sub-tasks, we use more public audio captioning datasets, including Clotho, AudioCaps[20], MACS[21], and part of WavCaps[22].

For CLAP and multi-task CLAP retrieval pre-training process, we use Clotho, AudioCaps, MACS and the entire Freesound part in WavCaps as our dataset. For CLAIP retrieval process, we use Clotho, AudioCaps, MACS and 40,000 Freesound clips, 50,000 audio_sl clips in WavCaps as our dataset.

For captioning pre-training process, we use Clotho, and the whole Freesound part in WavCaps as our dataset.

3.2. Text-to-audio Retrieval

For text-to-audio retrieval task, CLAP, CLAP-TAG, and CLAIP architecture are all trained with a pre-training process and a fine-tuning process. During both pre-training and fine-tuning stages, the retrieval model undergoes 20 epochs of training with a batch size of 128. The parameters of both the audio encoder and text encoder are initialized with pre-trained values, which warrants the use of a lower learning rate. For optimization, the Adam optimizer is employed. The learning rate undergoes a linear warm-up during the first epoch and subsequently follows a decay pattern using a cosine scheduler. In the pre-training phase, the maximum learning rate is set to 1×10^{-4}, while during fine-tuning, it is adjusted to 2×10^{-5}.

- System1: Ensemble the four top-performing models based on the mAP@10 metric on the evaluation set.
- System2: Ensemble the four top-performing models based on the R@10 metric on the evaluation set.
- System3: Ensemble the ten top-performing models based on the mAP@10 metric on the evaluation set.
- System4: Single model that performs the best on the mAP@10 metric.

3.3. Audio Captioning

For audio captioning task, we use audio encoders after audio-to-text retrieval as the pre-training feature extractors. The whole captioning model except the feature extractor is pre-trained for 10 epochs on the freesound and Clotho dataset, where the batch size is 32, Adam is used as the optimizer, and the learning rate is up to 5×10^{-4}. Then, we fine-tune the model for 15 epochs with the same settings on Clotho. During inference, beam search with a size of 3 is used. Different models are ensembled to further enhance the performance. Our submission setups are as follows:

- System1: Ensemble of the best model with multi-task training CLAP and the best model with original CLAP.
- System2: A single model with multi-task training CLAP.
- System3: Ensemble of six models with multi-task training CLAP, of which three adopt L1 loss and three adopt MSLE loss.
- System4: Ensemble of system3 and the best model with CLAIP.

4. RESULTS

4.1. Text-to-audio Retrieval

The performance of text-to-audio retrieval is presented in Table 1. System1 is an ensemble composed of the top-performing models, selected based on the mAP@10 metric on the development set. These four models are carefully chosen from the four best performing models of CLAIP, CLAP-TAG, and CLAP. The ensemble achieves an impressive result of 0.321 mAP@10. On the other hand, System2 is an ensemble of the top-performing models based on the R@10 metric on the development set. However, the obtained results are not satisfactory, indicating that the ensemble did not lead to improved performance. Moving on to System3, it comprises an ensemble of the ten top-performing models based on the mAP@10 metric on the development set. Surprisingly, it is observed that incorporating more models into the ensemble does not necessarily yield better effectiveness. Contrasting with the ensembles, System4 represents a single model that exhibits the best performance on the mAP@10 metric. This single model is fine-tuned from the pre-trained CLAIP model, showcasing exceptional performance with a peak achievement of 0.293 mAP@10 among single model.

4.2. Audio Captioning

The audio captioning performance is shown in and Table 2. As evident from the results, the incorporation of multi-task training into CLAP yields perform much better than baseline. And through ensembling different methods, the system achieves even better results across almost all evaluation metrics. The overall SPIDEr score reaches 0.305, while the SPIDEr-FL score reaches 0.294.

Submissions	R@1	R@5	R@10	mAP@10
Baseline	0.130	0.343	0.480	0.222
System1	**0.213**	**0.465**	**0.603**	**0.321**
System2	0.196	0.453	0.585	0.305
System3	0.206	0.461	0.600	0.316
System4	0.185	0.433	0.567	0.293

Table 1: Results for text-to-audio retrieval on Clotho evaluation set by submitted systems.

Submissions	MTR	CDEr	SPC	SPDr	SPDr-FL
Baseline	0.177	0.42	0.119	0.270	0.261
System1	0.191	0.471	**0.136**	0.304	**0.295**
System2	0.189	0.460	**0.136**	0.298	0.286
System3	0.190	0.468	0.135	0.302	0.292
System4	**0.192**	**0.474**	**0.136**	**0.305**	0.294

Table 2: Results for audio captioning by submitted systems. MTR, CDEr, SPC, SPDr, SPDr-FL denote METEOR, CIDEr, SPICE, SPIDEr and SPIDEr-FL, respectively

5. CONCLUSION

This paper proposes our submission to the DCASE 2023 challenge Task 6. Our approach utilizes a tri-encoder architecture, incorporating pre-trained encoders for audio, text, and image to establish relationships. Through multi-task training and fine-tuning, we achieve promising results, with a SPIDEr score of 0.305 and a SPIDEr-FL score of 0.294 for captioning, along with an mAP of 32.14 for text-to-audio retrieval.

6. ACKNOWLEDGMENT

We would like to thank Xuenan Xu, who provided valuable insights, assistance, and support throughout the entire research process.

7. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2017, pp. 374–378.

[2] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 830–834.

[3] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.

[4] X. Xu, Z. Zhang, Z. Zhou, P. Zhang, Z. Xie, M. Wu, and K. Q. Zhu, "Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data," *arXiv preprint arXiv:2303.07902*, 2023.

[5] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pre-training with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 905–909.

[8] X. Xu, M. Wu, and K. Yu, "A comprehensive survey of automated audio captioning," *arXiv preprint arXiv:2205.05357*, 2022.

[9] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," *arXiv preprint arXiv:2107.09817*, 2021.

[10] F. Gontier, R. Serizel, and C. Cerisara, "Automated audio captioning by fine-tuning bart with audioset tags," in *Detection and Classification of Acoustic Scenes and Events-DCASE 2021*, 2021.

[11] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022.

[12] E. Labbé, T. Pellegrini, and J. Pinquier, "Is my automatic audio captioning system so bad? spider-max: a metric to consider several caption candidates," *arXiv preprint arXiv:2211.08983*, 2022.

[13] X. Xu, Z. Xie, M. Wu, and K. Yu, "The SJTU system for DCASE2022 challenge task 6: Audio captioning with audio-text retrieval pre-training," DCASE2022 Challenge, Tech. Rep., July 2022.

[14] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[18] L. Beyer, P. Izmailov, A. Kolesnikov, M. Caron, S. Kornblith, X. Zhai, M. Minderer, M. Tschannen, I. Alabdulmohsin, and F. Pavetic, "Flexivit: One model for all patch sizes," *arXiv preprint arXiv:2212.08013*, 2022.

[19] 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.

[20] 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.

[21] I. Martin Morato and A. Mesaros, "Diversity and bias in audio captioning datasets," 2021.

[22] 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.