Dual-Encoder Audio Retrieval with PaSST and RoBERTa: A Contrastive and Distillation Approach

Technical Report

*Xichang Cai**

North China University of Technology
School of AI and Computer Science
Beijing 100144, China
caixc_ip@126.com

Weijie Luo, *Jiafeng Li*

North China University of Technology
School of AI and Computer Science
Beijing 100144, China
jiejiejie196@gmail.com

ABSTRACT

This technical report describes the Cai_NCUT team's submissions to the language-based audio retrieval task of the 2025 DCASE Challenge (Task 6). Our systems are built upon the dual encoder architecture, mapping audio clips and textual queries into a joint embedding space using pretrained backbones. In this submission, we explore the use of the recently proposed Time-Aware Spectrogram Transformer (PASST) as the audio encoder and RoBERTa as the text encoder. We introduce a two-stage training pipeline involving contrastive learning and a self-distillation phase to leverage cross-modal soft alignments. Our best single system, based on PASST and RoBERTa-large, achieves a mAP@16 of 0.32 on the ClothoV2 test split[1].

***Index Terms*—** dual encoder architecture, embedding space, PASST, RoBERTa, embedding space

1. INTRODUCTION

Language-based audio retrieval aims to retrieve relevant audio recordings from a database given a natural language query. This task has received increasing attention due to its practical value in multimedia search, audio indexing, and assistive technology[3]. Unlike traditional tag-based retrieval systems, language-based retrieval allows users to express complex auditory concepts — such as specific sound events, acoustic textures, or temporal dynamics — through free-form textual descriptions, enabling more flexible and intuitive access to sound data[2].

However, this task poses several challenges. First, audio and language are fundamentally different modalities, requiring models to learn meaningful cross-modal representations. Second, most datasets contain only positive audio-caption pairs, making it difficult to properly model semantic similarity for non-matching (but related) pairs[5]. Lastly, high-quality audio-text annotations are limited in size and diversity compared to purely textual corpora or image-text datasets.

To address these issues, recent work has adopted dual-encoder architectures, where audio and text are independently embedded into a shared space using pretrained models. The similarity between a caption and a candidate audio is computed as the distance between their respective embeddings. Training is typically conducted via contrastive learning on positive pairs, assuming all other combinations are negatives.

In this work, we follow this paradigm and introduce a dual-encoder system that employs PaSST as the audio encoder and RoBERTa-large as the text encoder. PaSST, a patch-based spectrogram transformer pretrained on AudioSet, offers strong temporal modeling and efficient representation of long audio segments. RoBERTa-large provides rich sentence-level semantics due to its large-scale pretraining on diverse text corpora.

To further improve retrieval performance, we propose a two-stage training pipeline that combines supervised contrastive learning with self-distillation. In the second stage, we use an ensemble of pretrained models to estimate soft alignment probabilities between all audio-caption pairs, providing richer supervision beyond binary labels.
Our method achieves strong performance on the ClothoV2 benchmark and demonstrates the effectiveness of combining strong backbone models with soft alignment strategies for language-based audio retrieval.

2. METHOD

Our system uses a dual-encoder architecture that maps audio and text inputs into a shared embedding space. The similarity between an audio clip aaa and a caption ccc is computed using cosine similarity between their L2-normalized embeddings. During training, we apply temperature-scaled contrastive loss to align matching pairs and separate non-matching ones:

$$L_{contrastive} = H(p_i , q_i)$$

To address semantic overlaps and false negatives, we introduce a self-distillation stage. An ensemble of pretrained models estimates soft alignment scores for all (ai,cj)(a_i, c_j)(ai,cj) pairs, which are used as soft targets in a distillation loss:

$$L_{distill} = H(\hat{p}_i , \hat{p}_i)$$

The final loss combines both objectives to enhance retrieval accuracy and robustness.

3. DATASETS

We utilized three widely used audio-text paired datasets in this work: **ClothoV2**, **AudioCaps**. These datasets differ in source, annotation quality, and linguistic style, making them complementary and beneficial for improving the robustness and generalization of our retrieval model. All three were used during the pretraining stage, while only **ClothoV2** was used during the knowledge distillation fine-tuning stage to match the evaluation setting of the DCASE 2025 challenge.

ClothoV2 is a high-quality audio-text dataset designed for audio captioning and retrieval tasks, provided by Tampere University. It consists of 5930 environmental sound clips, each lasting between 10 and 30 seconds and annotated with five human-written English captions. Each caption contains between 8 and 20 words and is typically detailed and acoustically grounded, such as: “a person is walking along a gravel path[6].”

We adopted the official data split: 3840 for training, 1045 for validation, and 1045 for testing. The evaluation set used for leaderboard scoring contains 1000 unseen audio-caption pairs. As ClothoV2 offers high-quality, semantically accurate annotations, we used it exclusively during the fine-tuning phase for knowledge distillation.
During training, we included all five captions for each audio clip as positive pairs to increase data diversity and training stability.

AudioCaps is a large-scale paired dataset derived from AudioSet, containing approximately 51,000 samples. Each sample consists of a 10-second audio clip and a single human-written caption. The average caption length is 9.8 words. The descriptions are typically concise and focused on a single acoustic event (e.g., “a man is speaking outdoors”)[8].

We merged the training, validation, and test splits of AudioCaps into a single large dataset for use in the pretraining stage only.

4. PREPROCESSING

To match the input requirements of the **PaSST** model, we applied a consistent preprocessing pipeline across all audio data:

- **Sample rate**: All audio clips were resampled to 32 kHz.
- **Segment handling**: Clips longer than 30 seconds were randomly cropped, while shorter clips were zero-padded.
- **Spectrogram extraction**: We used the log-Mel spectrogram configuration recommended by the original PaSST implementation.
- **Chunking**: Since PaSST supports input lengths up to 10 seconds, each audio clip was split into non-overlapping 10-second segments. Embeddings were extracted per segment and then averaged.

For text preprocessing:

- All captions were lowercased and stripped of punctuation.
- Tokenization was performed using the WordPiece tokenizer associated with RoBERTa.
- Captions were truncated or padded to a maximum of 32 tokens.

In ClothoV2, it is common to find multiple audio clips sharing the same caption. Rather than removing such duplicates, we preserved them during training to help the model learn semantic similarity and partial matches between audio and text.

5. MODELS

PASST: In our system, we employ the Patchout Spectrogram Transformer (PaSST) as the sole audio encoder. PaSST is a transformer-based model designed for efficient audio tagging and retrieval. It adapts the Vision Transformer (ViT) architecture to log-Mel spectrogram inputs by dividing them into 2D patches, which are then processed through a self-attention network.
To reduce memory usage and improve efficiency, PaSST introduces patchout, a structured dropout strategy that randomly removes a subset of time and frequency patches during training[9]. This regularization improves generalization without significantly sacrificing performance.

We use the pretrained passt_s_p16_s16_128_ap468 variant, which contains 86.2 million parameters and achieves a mAP of 46.8 on AudioSet. As PaSST is designed for 10-second inputs, all longer audio clips are split into non-overlapping 10-second segments. Embeddings from each segment are extracted and averaged to obtain a fixed-length audio representation.
RoBERTa-large: We use **RoBERTa-large** as our text encoder, a transformer-based language model known for its strong sentence-level understanding capabilities. RoBERTa improves upon BERT by removing the next sentence prediction task and using larger pretraining corpora and longer training durations. It was pretrained on a 160GB English text corpus including BookCorpus and Wikipedia[7].

We adopt the HuggingFace version of RoBERTa-large with **354 million parameters**. During preprocessing, all captions are lowercased, punctuation is removed, and the text is tokenized using the WordPiece tokenizer. Each caption is truncated or padded to a maximum of **32 tokens** to enable batched processing.

To obtain a sentence-level embedding, we extract the [CLS] token from the final transformer layer, which is then projected into the shared embedding space using a linear layer. This global embedding captures the semantic content of the entire caption and enables effective matching with audio representations.

RoBERTa-large was selected due to its strong performance in prior language-based retrieval systems. Its deep architecture and rich pretraining allow for robust handling of diverse, descriptive audio captions across datasets.

6. EXPERIMENT

To evaluate the effectiveness of our dual-encoder model using PaSST and RoBERTa-large, we conducted experiments using two datasets: ClothoV2 and AudioCaps.
While many prior systems rely on larger combined training corpora (e.g., including WavCaps or synthetic captions), we intentionally limited the training data to high-quality human-annotated sources in order to isolate the contribution of the model architecture and training strategy.
The results are summarized in Table 1. Even when trained on only two datasets, our model achieved a mAP@10 of 37.21, significantly outperforming typical strong baselines under similar or larger-scale training conditions. In particular, we observed consistent improvements across recall metrics, with R@10 reaching 60.48 and R@5 at 47.45, demonstrating robust retrieval performance.

Trained on	mAP@10 (new annotations)	Map@10	R@1	R@5	R@10
ClothoV2	31.95	27.57	16.95	42.73	58.04
ClothoV2	32.54	28.20	16.44	43.89	58.89
Clotho、AudioCaps	35.21	30.97	19.56	46.45	59.48
Clotho、AudioCaps	37.21	32.75	19.80	47.45	60.48

Table 1: Retrieval performance (mAP@10, R@1, R@5, R@10) of different models trained on Clotho and AudioCaps.*The 1st and 3rd rows correspond to baseline models using the same data; the 2nd and 4th rows show the performance of our model trained under the same conditions. Our method achieves consistently better performance across all metrics.*

7. CONCLUSIONS

In this technical report, we presented a language-based audio retrieval system based on a dual-encoder architecture, employing **PaSST** as the audio encoder and **RoBERTa-large** as the sentence encoder. The model was pretrained on three complementary datasets — **ClothoV2**, **AudioCaps** — using contrastive learning to align audio and textual representations within a shared embedding space.

To enhance performance, we introduced a **two-stage training strategy**: the first stage involved supervised pretraining with binary matching assumptions; the second stage incorporated **knowledge distillation**, where soft alignment scores between audio-caption pairs were estimated from an ensemble and used as targets for fine-tuning. This approach allowed the model to better capture semantic overlaps and partial correspondences between modalities.

Our best single system, using only PaSST and RoBERTa, achieved a **mean Average Precision at** top 16 of 31.6 on the **ClothoV2** test split — a strong result that outperforms many prior systems from previous DCASE challenges, without relying on metadata or synthetic text augmentation.

The results demonstrate that even with a single audio encoder, robust retrieval performance can be achieved through careful model selection and training design. In future work, we plan to explore the use of alternative audio encoders, incorporate metadata and large language model-generated captions, and develop fusion-based ensemble strategies to further push performance.

8. REFERENCES

[1] http://dcase.community/workshop2025/.

[2] P. Primus, G. Widmer, “AKNOWLEDGEDISTILLATION APPROACHTOIMPROVING

[3] LANGUAGE-BASEDAUDIORETRIEVALMODELS,” in Proc. of the 8th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE, Helsinki, Finland, 2024.

[4] S. Lou, X. Xu, M. Wu, and K. Yu, “Audio-text retrieval in context,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Singapore, 2022.

[5] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, “On metric learning for audio-text cross-modal retrieval,” in 23rd Annual Conf. of the Int. Speech Communication Association, Interspeech, Incheon, Korea, 2022.

[6] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. *Clotho: an audio captioning dataset.* In Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP), 736–740. 2020.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” CoRR, vol. abs/1907.11692, 2019.

[8] C. D. Kim, B. Kim, H. Lee, and G. Kim, “AudioCaps: Gen erating captions for audios in the wild,” in Proc. of the North American Ch. of the Ass. for Computational Linguistics: Hu man Language Technologies, NAACL-HLT, 2019.

[9] K.Koutini, J. Schl¨ uter, H. Eghbal-zadeh, and G. Widmer, “Ef f icient training of audio transformers with patchout,” in 23rd Annual Conf. of the Int. Speech Communication Association, Interspeech, 2022.

[10] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: To wards story-like visual explanations by watching movies and reading books,” in IEEE Int. Conf. on Computer Vision, ICCV, 2015.