A CROSS-MODAL ATTENTION APPROACH TO LANGUAGE-BASED AUDIO RETRIEVAL

Technical Report

Óscar Calvet, Doroteo T. Toledano

Audio, Data Intelligence and Speech research group (AUDIAS)
Escuela Politecnica Superior
Universidad Autonoma de Madrid
oscar.calvet@estudiante.uam.es

ABSTRACT

This report presents the systems we developed for the 2025 DCASE Language-Based Audio Retrieval challenge (task 6). We use a bi-encoder architecture and propose a novel cross modal attention approach in order to calculate the similarity between the embeddings produced by both models. We make use of pretrained encoders in both modalities: PaSST is used for encoding audio and RoBERTa for encoding text. We trained our system on WavCaps, AudioCaps, ClothoV2 and TACOS using contrastive learning. The best single system that we were able to produce reaches a mAP@10 of 38.293 on the ClothoV2 test split and a mAP@16 of 44.203 using the task specific improved notations. An ensemble of the models presented archieves a mAP@10 of 40.423 on the ClothoV2 test split and a mAP@16 of 46.864 using the task specific improved notations.

***Index Terms*—** Language-based audio retrieval, Audio transformer, Cross modal attention

1. INTRODUCTION

Task 6 of the 2025 DCASE Challenge once again invites participants to build systems that retrieve audio recordings from a large database given a free-form textual description. These language-based retrieval systems are attractive in practice because they let users express arbitrary acoustic concepts, ranging from concrete events to more abstract auditory scenes, without being confined to a fixed taxonomy of tags. Successfully matching a raw waveform to a sentence, however, remains technically demanding: the model must learn a shared representation in which distances between heterogeneous modalities (audio and text) meaningfully reflect semantic similarity.

This year's edition introduces a substantial update to the benchmark material. In addition to the familiar Clotho V2 training split, new human-verified relevance annotations for the entire development-testing set are provided. Concretely, the dev-test split contains 1069 natural-language queries (one per caption) and their corresponding audio files; for every query, several audio recordings are explicitly marked as relevant. These many-to-many relevance labels allow participants to evaluate retrieval quality with finer-grained metrics. Therefore, the 2025 edition introduces a notable expansion of the benchmark material.

In the standard practice, a dual-encoder is trained with a contrastive loss so that a text caption and its paired waveform are projected to nearby points in a joint embedding space. At inference time, relevance is scored with the cosine similarity between the global audio and text embeddings, and audio clips are ranked accordingly. While this single-vector, cosine-based matching has proved effective, it compresses each modality into one vector and therefore discards fine-grained structure—e.g., the alignment between individual words and short acoustic events. Recent studies on audio-text retrieval indicate that explicitly modeling token-level interactions with cross-modal attention can capture richer semantic cues and improve ranking performance [1, 2]. In this report, we adopt a new supervision regime and investigate how richer relevance signals can be leveraged to improve retrieval effectiveness.

2. CROSS MODAL ATTENTION IN TEXT TO AUDIO RETRIEVAL

We present a cross-modal attention-based method for text-to-audio retrieval that aims to effectively align textual and audio representations based on already existing studies [3].

Given textual embeddings $T \in \mathbb{R}^{N \times d_t}$ and audio embeddings $A \in \mathbb{R}^{M \times d_a}$ where N denotes the number of embeddings of a sentence, d_t the dimension of the text embeddings, M the number of embeddings of an audio and d_a the dimension of the audio embeddings, we first project them into a common space of dimension d. Specifically, textual embeddings T are projected through a linear transformation to obtain the queries vector (Q):

$$Q = W_q T \quad Q \in \mathbb{R}^{N \times d} \tag{1}$$

where $W_q \in \mathbb{R}^{d \times d_t}$ are learned parameters.

Audio embeddings are projected independently using two separate linear transformations to generate key (K) and value (V) vectors:

$$K = W_k A, \quad K \in \mathbb{R}^{M \times d} \tag{2}$$

$$V = W_v A, \quad V \in \mathbb{R}^{M \times d} \tag{3}$$

where $W_k, W_v \in \mathbb{R}^{d \times d_a}$ are also trainable parameters.

Next, we apply multi-head attention, defined as:

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{4}$$

where h indicates the number of heads, $W^O \in \mathbb{R}^{d\times d}$ are trainable parameters and each attention head is implemented following the standard definition [4]:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

where W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d\times d_h}$ are all trainable parameters and d_h denotes the head size.

The output of the attention module is passed through another linear layer:

$$H = W_o \cdot \text{MultiHead}(Q, K, V), \quad H \in \mathbb{R}^{N\times d} \tag{6}$$

with parameters $W_o \in \mathbb{R}^{d\times d}$ also being trainable.

Finally, we compute the similarity between these refined cross-modal representations H and the original textual embeddings T using the average cosine distance between them as follows:

$$\text{sim}(T, H) = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i \cdot H_i}{|T_i||H_i|} \tag{7}$$

To train our models we rely on the normalized temperature cross-entropy loss [5], which transforms the similarities into conditional probabilities using a temperature-scaled softmax.

3. EXPERIMENTAL SETUP

3.1. Datasets

Our training process was executed utilizing several datasets: ClothoV2 [6], AudioCaps [7], WavCaps [8] and TACOS [9].

3.1.1. ClothoV2

ClothoV2 [6] contains audio clips ranging from 10 to 30 seconds in duration, each accompanied by five descriptive captions of lengths between 8 and 20 words. Following the dataset organizers' recommended splits, the development set includes 3840 training, 1045 validation, and 1045 testing audio clips. Performance was monitored on the validation split, while test split results are reported here and with the extended notations introduced this year.

3.1.2. AudioCaps

AudioCaps [7] consists of 51,308 audio clips sourced from AudioSet [10], each paired with a single human-written caption averaging 9.8 words. The training, validation, and test splits were combined into a unified dataset utilized during the pretraining stage.

3.1.3. WavCaps

WavCaps [8] provides weakly-labeled synthetic captions for 403,050 audio clips from multiple sources, including FreeSound, BBC Sound Effects, SoundBible, and AudioSet. Captions, generated via GPT3.5-turbo, average 7.8 words each. To ensure compliance with current guidelines, overlapping data with Clotho evaluation sets were excluded.

3.1.4. TACOS

TACOS (Temporally-Aligned Audio CaptiOnS) [9] is a dataset designed to facilitate fine-grained audio-language modeling. It comprises 12358 real-world audio recordings sourced from Freesound, each annotated with multiple free-text captions that are temporally aligned to specific regions within the audio. In total, the dataset includes 47,748 such region-level annotations, amounting to approximately 98 hours of labeled audio and covering over 76.6 hours of unique content. On average, each audio clip contains 3.57 annotated regions. TACOS introduces strong supervision for contrastive audio-text learning by aligning each caption with its respective time segment, thus enabling frame-wise contrastive training. Nevertheless, in our experiments we will only use the weak annotations which do not include timestamps.

3.2. Audio Embedding Models

We only employed a single audio embedding architecture, PaSST [11], which leverages pretrained vision transformer parameters, fine-tuned on AudioSet, and employs a patch-dropping strategy for computational efficiency. It produces an audio embedding every 10 seconds.

3.3. Sentence Embedding Model

We employed RoBERTa-large [12] as the sentence embedding model due to its strong performance in previous retrieval tasks. RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a transformer-based encoder that improves upon BERT by training with larger mini-batches, longer sequences, and dynamic masking. It omits the next sentence prediction objective used in BERT, focusing instead on masked language modeling with more aggressive training settings.

RoBERTa-large consists of 24 transformer layers, 16 attention heads per layer, and 355 million parameters in total. In our setup, we utilized the output corresponding to the final transformer layer as the sentence representation, which includes a summary embedding and the embeddings of all the tokens in the phrase.

3.4. Preprocessing

Audio data preprocessing involved extracting random 30-second segments for clips exceeding this length or zero-padding shorter clips to the batch's maximum length. Text data were standardized to a lowercase and punctuation-free format, tokenized with RoBERTa tokenizer, and padded or truncated to a maximum length of 32 tokens.

3.5. Training Procedure

We trained three different configurations consisting of the task baseline architecture (which we will refer to from now on as PaSST–RoBERTa Base), a cross attention system which only used the sentence embedding token from RoBERTa, and finally another cross attention system which utilized all of the text embeddings. We used 8 attention heads and a joint embedding space of dimension 1024.

We trained both encoders and the cross-modal attention heads on combined datasets (AudioCaps, WavCaps, ClothoV2, TACOS

	Clotho test split				Improved captions	
Model	mAP@10	R@1	R@5	R@10	mAP@16	mAP@10
PaSST–RoBERTa Base	37.389	24.976	54.086	68.172	42.324	39.664
Sentence embedding	35.769	23.655	52.766	66.258	42.434	39.827
Full text embeddings	35.77	23.828	52.057	65.742	41.375	38.908

Table 1: Performace for the systems after pre-training

	Clotho test split				Improved captions	
Model	mAP@10	R@1	R@5	R@10	mAP@16	mAP@10
PaSST–RoBERTa Base	38.293	25.263	56	69.282	44.203	41.662
Sentence embedding	37.495	24.784	54.947	68.650	43.926	41.332
Full text embeddings	37.901	25.818	54.43	67,617	42.457	40.061
Ensemble	40.423	27.732	58.201	71.732	46.864	44.176

Table 2: Performace for the systems after knowledge distillation process

weak) employing the Adam optimizer with batch sizes of 128 for the PaSST–RoBERTa Base and sentence embedding systems and 96 for the full text embeddings model. After one warm-up epoch, a cosine annealing schedule was used to decrease the learning rate from 2×10^{-5} to 10^{-7} over 10 epochs.

After the training, fine-tuning with knowledge distillation was executed in the same style as in [13] on ClothoV2, AudioCaps and TACOS weak by applying the same training setup. Predictions from three models were averaged to estimate caption-audio correspondences. All experiments were performed with a temperature parameter $\tau = 0.05$ and loss balancing factor $\lambda = 1$.

4. RESULTS

We evaluated the performance of three system configurations: the baseline architecture PaSST–RoBERTa Base model, a cross-attention model using only the sentence embedding from RoBERTa, and a cross-attention model leveraging the full set of token-level embeddings. Additionally, we evaluated an ensemble of these models.

Table 1 summarizes results obtained after the initial pretraining phase. While all systems show comparable performance, the baseline model slightly outperforms the attention-based models in mAP@10 on the ClothoV2 test set.

Table 2 summarizes results after the knowledge distillation phase. The figures indicate that, although modelling fine-grained token interactions can capture richer audio-language correspondence, a single-vector matching strategy remains highly competitive when both encoders are extensively pre-trained on diverse audio-caption corpora. Our best performing model obtains a 38.293 mAP@10 score of the Clotho test split and a 44.203 mAP@16 score on the improved captions.

Finally, the ensemble of all three systems further improves performance archieving a mAP@10 of 40.423 and a mAP@16 of 46.864 and consistent gains across recall metrics. These gains confirm that the three systems capture complementary cues: the dual-encoder contributes strong global alignments, while the attention-based models compensate with finer local correspondences.

5. REFERENCES

[1] Q. Wang, J.-C. Gu, and Z.-H. Ling, "Multiscale matching driven by cross-modal similarity consistency for audio-text retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2403.10146

[2] Y. Xin and Y. Zou, "Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions," 2025. [Online]. Available: https://arxiv.org/abs/2307.15344

[3] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," 2023. [Online]. Available: https://arxiv.org/abs/2303.05681

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[6] 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.

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

[8] 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.

[9] P. Primus, F. Schmid, and G. Widmer, "Tacos: Temporally-aligned audio captions for language-audio pretraining," 2025. [Online]. Available: https://arxiv.org/abs/2505.07609

[10] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[11] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022, 23rd Annual Conference of the*

International Speech Communication Association, Incheon, Korea, 18-22 September 2022. ISCA, 2022, pp. 2753–2757. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-227

[12] 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," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[13] P. Primus, F. Schmid, and G. Widmer, "Estimated audio-caption correspondences improve language-based audio retrieval," 2024. [Online]. Available: https://arxiv.org/abs/2408.11641