ENHANCED AUDIO-TEXT RETRIEVAL WITH MULTI-POSITIVE LEARNING AND HARD NEGATIVE MINING

Technical Report

Saubhagya Pandey saubhagya23@iitk.ac.in
Khushal Wadhwa kwadhwa23@iitk.ac.in
Ayush Goyal gayush23@iitk.ac.in
Kanak Khandelwal kanakk23@iitk.ac.in

Indian Institute of Technology Kanpur, India

ABSTRACT

This paper presents an enhanced implementation of audio-text cross-modal retrieval for DCASE 2025 Task 6, featuring advanced contrastive learning techniques. Our system implements a dual-encoder architecture using PaSST (Patch-out Fast Spectrogram Transformer) for audio encoding and RoBERTa-large for text encoding, enhanced with multi-positive learning and hard negative mining strategies. The proposed method introduces progressive training with staged activation of advanced techniques, achieving significant performance improvements over baseline approaches. Experimental evaluation on the Clotho dataset demonstrates competitive retrieval performance with R@1 of 18.68%, R@5 of 44.77%, R@10 of 59.35%, and mAP@10 of 30.01%. The implementation supports mixed precision training and comprehensive evaluation metrics for robust cross-modal retrieval.

***Index Terms*—** Audio-text retrieval, contrastive learning, multi-positive learning, hard negative mining, cross-modal retrieval, PaSST, RoBERTa

1. INTRODUCTION

Language-based audio retrieval has emerged as a critical task in multimedia understanding, enabling users to search through audio collections using natural language descriptions. The DCASE 2025 Task 6 focuses on this challenging problem, requiring systems to effectively bridge the semantic gap between audio content and textual descriptions.

Recent advances in contrastive learning have shown promising results for cross-modal retrieval tasks. However, traditional contrastive learning approaches often treat all negative samples equally and may miss opportunities to leverage semantic relationships between samples. This work addresses these limitations by introducing enhanced contrastive learning techniques including multi-positive learning and hard negative mining.

Our contributions include: (1) Implementation of multi-positive learning that identifies semantically similar samples as soft positives with weighted loss computation, (2) Hard negative mining strategies that focus training on challenging negative samples, (3) Progressive training approach with staged activation of advanced techniques, and (4) Comprehensive evaluation framework supporting multiple datasets and metrics.

2. METHODOLOGY

2.1. Model Architecture

Our system implements a dual-encoder architecture for cross-modal audio-text retrieval, consisting of separate encoders for audio and text modalities that are trained jointly using enhanced contrastive learning.

2.1.1. Audio Encoder

The audio encoder utilizes PaSST (Patch-out Fast Spectrogram Transformer), a state-of-the-art audio transformer architecture. For handling variable-length audio inputs, we implement a segment-based processing approach:

- Long audio files (more than 10s) are split into overlapping 10-second segments
- Each segment is processed through PaSST with configurable patch dropout (temporal: 15, frequency: 2)
- Duration-based aggregation combines segment embeddings using mean pooling
- A linear projection layer maps features to 1024-dimensional embedding space

The audio processing pipeline includes mel-spectrogram extraction with the following parameters: 128 mel-bands, 32kHz sampling rate, 800-sample window length, and 320-sample hop size.

2.1.2. Text Encoder

The text encoder employs RoBERTa-large for contextual text understanding:

- Text preprocessing includes lowercase conversion and punctuation removal
- RoBERTa tokenizer with 32-token maximum length handles variable-length captions
- Contextualized embeddings are extracted from the [CLS] token representation
- A linear projection layer maps to the same 1024-dimensional space as audio

Both audio and text embeddings are L2-normalized before similarity computation to ensure stable training dynamics.

2.2. Enhanced Contrastive Learning

2.2.1. Multi-Positive Learning

Traditional contrastive learning treats sample pairs as either positive (matched) or negative (unmatched). Our multi-positive learning approach extends this by identifying semantically similar samples as soft positives:

$$\mathcal{L}_{mp} = -\log \frac{\sum_j w_{ij} \exp(s_{ij}/\tau)}{\sum_j w_{ij} \exp(s_{ij}/\tau) + \sum_{k \in \mathcal{N}} \exp(s_{ik}/\tau)} \quad (1)$$

where w_{ij} represents positive weights (1.0 for hard positives, 0.3 for soft positives), s_{ij} is the cosine similarity, τ is the learnable temperature parameter, and $\mathcal{N}$ is the negative set.

Soft positives are identified using intra-modal similarity thresholds (0.75 for text-text and audio-audio similarities), capturing semantic relationships beyond exact matches.

2.2.2. Hard Negative Mining

Hard negative mining focuses training on challenging negative samples that are most likely to be confused with positives. We implement three mining strategies:

- **Hardest**: Select negatives with highest similarity to positives
- **Semi-hard**: Select negatives within a margin of positive similarity
- **Random**: Baseline random negative sampling for comparison

The number of hard negatives per positive pair is configurable (default: 5), with a margin parameter (0.15) controlling semi-hard selection.

2.2.3. Progressive Training

To prevent optimization interference, advanced techniques are activated progressively:

- Epochs 0-7: Standard InfoNCE contrastive loss
- Epoch 8+: Multi-positive learning activation
- Epoch 10+: Hard negative mining activation

This staged approach allows the model to learn basic cross-modal alignments before introducing more sophisticated loss components.

3. EXPERIMENTAL SETUP

3.1. Datasets

The primary evaluation dataset is Clotho v2.1, which provides high-quality audio-caption pairs:

- Training: Development split (3,839 audio files)
- Validation: Validation split (1,045 audio files)
- Testing: Evaluation split (1,045 audio files)

The system supports augmentation with AudioCaps and WavCaps datasets for larger-scale training, though computational constraints limited this evaluation to Clotho only.

Table 1: Text-to-audio retrieval performance on Clotho evaluation set

Metric	Performance (%)
R@1	18.68
R@5	44.77
R@10	59.35
mAP@10	30.01
mAP@16 (multiple positives)	34.68

3.2. Training Configuration

Training employed the following configuration:

- Hardware: NVIDIA L4 GPU (24GB memory)
- Batch size: 24 (both training and evaluation)
- Learning rate: Linear warmup (1 epoch) + cosine decay (15 epochs)
- Peak learning rate: 2e-5, minimum: 1e-7
- Mixed precision: 16-bit automatic mixed precision
- Total epochs: 20
- Optimizer: AdamW without weight decay

The learnable temperature parameter was initialized to 0.05 and remained trainable throughout training.

3.3. Evaluation Metrics

Standard cross-modal retrieval metrics were employed:

- Recall at K (R@1, R@5, R@10): Fraction of queries with relevant item in top-K
- Mean Average Precision at 10 (mAP@10): Average of precision scores
- mAP@16 with multiple positives: Extended evaluation considering multiple relevant items per query

4. RESULTS

4.1. Retrieval Performance

Table 1 presents the text-to-audio retrieval performance on the Clotho evaluation set.

The results demonstrate competitive performance for the enhanced contrastive learning approach. The improvement in mAP@16 over mAP@10 (34.68% vs 30.01%) indicates the system's ability to identify multiple relevant audio files for given text queries.

4.2. Training Dynamics

The progressive training approach showed stable convergence with staged activation of advanced techniques. The learnable temperature parameter converged to approximately 0.04, indicating optimal scaling for the similarity matrix. Multi-positive learning activation at epoch 8 and hard negative mining at epoch 10 demonstrated smooth integration without training instability.

5. IMPLEMENTATION DETAILS

The system is implemented using PyTorch Lightning for distributed training support and includes several engineering optimizations:

- Custom audio loading with efficient 30-second segment extraction using FFmpeg
- Duration-based audio padding and subsampling for consistent input shapes
- Custom batch collation for handling variable-length sequences
- Torch.compile() optimization for compatible GPUs (SM 7.0+)
- Comprehensive dataset filtering to exclude corrupted and forbidden files

The modular design supports easy extension to additional datasets and alternative encoder architectures.

6. FUTURE DIRECTIONS AND ONGOING WORK

6.1. Compositional Audio-Text Retrieval Framework

Building upon the enhanced contrastive learning approach presented in this work, we explored the adaptation of compositional text-to-image retrieval methodologies for audio-text retrieval tasks. Inspired by the Cola benchmark framework for compositional understanding, we investigated a two-stage retrieval mechanism that leverages both our base contrastive model and multimodal adapter components.

The proposed compositional framework operates through a hierarchical retrieval pipeline:

- **Stage 1 - Semantic Retrieval**: The base model with multi-positive learning and hard negative mining generates initial candidate rankings based on learned audio-text embeddings
- **Stage 2 - Compositional Reranking**: A multimodal adapter performs fine-grained reranking of top-N candidates, focusing on compositional understanding of complex audio scene descriptions

This approach aims to address limitations in current audio-text retrieval systems when handling compositional queries that describe multiple sound sources, temporal relationships, or complex acoustic scenes. The multimodal adapter, implemented as a transformer-based cross-attention mechanism, was designed to capture fine-grained correspondences between textual compositional elements and audio spectral-temporal patterns.

6.1.1. Implementation Strategy

The compositional framework builds upon our validated base architecture by introducing a specialized reranking module. The base model, trained with enhanced contrastive learning techniques described in Section 2.2, provides robust initial retrieval performance. The multimodal adapter then refines these predictions by analyzing compositional relationships within the top-50 retrieved candidates.

Initial experiments demonstrated promising directions for improved handling of complex audio scene descriptions. However, the full implementation and evaluation of this compositional framework remains incomplete due to computational constraints and project timeline limitations.

6.1.2. Expected Contributions

The completed compositional framework is expected to provide several key improvements:

- Enhanced retrieval accuracy for complex, multi-element audio scene descriptions
- Better handling of temporal and spatial relationships in audio-text correspondences
- Improved generalization to compositional queries not seen during training
- Systematic evaluation methodology for compositional audio-text understanding

This work represents a natural extension of the current approach and will be pursued in future research to advance the state-of-the-art in compositional audio-text retrieval.

7. CONCLUSION

This work presents an enhanced audio-text retrieval system featuring multi-positive learning and hard negative mining within a progressive training framework. The approach demonstrates competitive performance on the Clotho dataset while providing a flexible foundation for cross-modal retrieval research.

Key contributions include the systematic integration of advanced contrastive learning techniques, progressive training methodology, and comprehensive evaluation framework. The staged activation approach successfully prevents optimization interference while enabling sophisticated loss formulations.

Future work directions include scaling to larger multi-dataset training, exploring alternative mining strategies, and investigating attention-based cross-modal fusion mechanisms. The open-source implementation facilitates reproducibility and further research in cross-modal retrieval.

8. ACKNOWLEDGMENT

We thank Prof. Vipul Arora from the Department of Electrical Engineering, Indian Institute of Technology Kanpur, for his valuable guidance and mentorship throughout this project. We acknowledge the Indian Institute of Technology Kanpur for providing the computational resources used in this research. We also acknowledge the original DCASE 2025 Task 6 baseline implementation that served as the foundation for this enhanced approach.

9. REFERENCES

[1] P. Primus, F. Schmid, and G. Widmer, "Estimated audio–caption correspondences improve language-based audio retrieval," in *Proc. Detection and Classification of Acoustic Scenes and Events 2024 Workshop*, Tokyo, Japan, 2024, pp. 121–125.

[2] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech 2022*, 2022, pp. 2753–2757.

[3] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” in *Proc. IEEE ICASSP*, 2020, pp. 736–740.

[5] C. D. Kim, B. Kim, H. Lee, and G. Kim, “AudioCaps: Generating captions for audios in the wild,” in *Proc. NAACL-HLT*, 2019, pp. 119–132.

[6] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, ”WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” arXiv, no. 2303.17395, 2023.

[7] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[8] A. Ray, F. Radenovic, A. Dubey, B. Plummer, R. Krishna, and K. Saenko, “Cola: A benchmark for compositional text-to-image retrieval,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 46433–46445.

[9] DCASE 2025 Challenge, `http://dcase.community/challenge2025/`.