2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

# Contrastive Learning Scheme for Audio-Text Classification using Contextual Information

Yao Xiao[1], Shiheng Zhang[1], Feiyang Xiao[1], Qiaoxi Zhu[2], and Jian Guan[1*]
[1]Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology,
Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia

***Abstract*—This report describes my approach to the challenges of semi-supervised urban acoustic scene and time-aware classification. I constructed an audio classification model based on contrastive learning by combining audio and text modal training. This model integrates an audio encoder based on SE-ResNet-Transformer and a text encoder based on BERT. The model is trained using a hybrid loss function that combines supervised classification and comparison objectives, aligning multimodal features in a shared semantic space. Experiments demonstrate that this model improves the representational power of audio features through multimodal learning and achieves good performance.**

## I. Introduction

The city and time-aware semi-supervised acoustic scene classification is an extension of semi-supervised acoustic scene classification with domain shift, addressing the challenge of generalization across different cities. The challenge explicitly provides city-level location and timestamp metadata for each audio sample, but audio data often contains a large amount of unlabeled data and only a limited number of labeled examples, which is consistent with real-world scenarios [1].

Relying solely on audio data to train a model not only wastes audio metadata but also suffers from poor performance when used on datasets containing large amounts of unlabeled data. In such cases, the model needs to be able to fully utilize the metadata of limited labeled examples and samples, as well as the large amount of unlabeled data, to improve model generalization. To leverage both temporal and location metadata, which are equally important, consider converting them into text and training them alongside the audio tags.

This technical report introduces a lightweight audio-text contrastive learning integration method based on Contrastive Language-Audio Pretraining (CLAP) [2]. Analysis of the development set reveals that while labeled audio provides limited information, all data includes city-level locations and timestamps, which are crucial for classifying audio. To fully leverage this location and time metadata,I developed an audio-text contrastive learning method based on the CLAP model, incorporating both location and timestamps as text into the model output, thus enabling the model to incorporate context. Furthermore, to leverage unlabeled data, the model is first trained on labeled data, then the trained model is used to predict and assign pseudo-labels to the unlabeled data. A second training run uses all the data for the final model.

## II. Proposed Systems

The model I proposed is a multimodal framework with context-enhanced capabilities. Compared with the original CLAP model, it adds contextual information such as city location and timestamp to the text input.

### *A. Audio Encoder*

The model contains three residual blocks in terms of audio processing, each of which is composed of two convolutional layers, and Squeeze-and-Excitation (SE)[3] is added to construct information-rich features, and the number of channels of the three residual blocks is 64, 128, 256, respectively, and the kernel size of each volume sublayer is 3×3 through global draw pooling and two fully connected layers to learn channel attention weights. After convolutional extraction, an adaptive average pooling layer is used. For the Transformer encoder, I set the number of heads to 4, the number of layers to 2, and the input dimension to 256. The final output is pooled evenly over the time dimension to obtain a global audio representation that is mapped to the final embedding dimension through a linear layer.

### *B. Text Encoder*

The model uses the pre-trained BERT model for text processing and the HuggingFace bert-base-uncased model[4] as the text encoder. The parameters of the first six layers of the BERT encoder are frozen, and the output vector is processed through a linear layer, and LayerNorm is mapped to a vector of the same dimension as the audio.

### *C. Joint Loss*

To enable simultaneous training of audio and text modalities, the model integrates the audio encoder, text encoder, a learnable temperature parameter, and an audio classifier composed of two Linear layers with a GELU activation. The classifier maps audio embeddings to specific scene labels for computing the classification loss. During training, both audio and text embeddings participate in contrastive learning, where the temperature parameter controls the similarity scaling for contrastive loss computation. The classification loss is derived from the audio embeddings and their corresponding ground-truth labels, while the contrastive loss is computed from paired audio–text embeddings. The total loss is dominated by the classification loss, with the contrastive loss serving as an auxiliary objective. To enhance pseudo-label utilization and improve training performance, confidence scores are assigned to each predicted label during inference, and only high-confidence pseudo-labels are retained for subsequent training.

## III. EXPERIMENTS

### A. Experimental Setup

To extract audio features, all recordings were resampled to 44.1 kHz. A short-time Fourier transform (STFT) was applied using a Hanning window with a length of 40 ms and a hop size of 20 ms, with an FFT size of 2048. The magnitude spectrum was processed using 64 Mel filter banks, followed by a logarithmic transformation, resulting in log-Mel spectrograms of size 500 × 64, where 500 corresponds to the number of time frames and 64 to the frequency bins. These log-Mel spectrograms served as the audio input features for the model. For text features, the BERT tokenizer was used to generate token ID sequences and corresponding attention masks. Each token ID sequence was represented as a fixed-length vector of 64 integer indices. During training, the Adam optimizer was employed with a learning rate of 0.001, a batch size of 8, and a cosine annealing learning rate schedule.

### B. Result

I trained my model in the development set of the APSIPA ASC 2025 Challenge. After training on labeled data, the model was able to predict pseudo-labeled data with a confidence level above 0.9 in most cases. After retraining on pseudo-labeled data, the validation set accuracy reached 0.9494, a slight improvement over the baseline system[1].

## IV. CONCLUSION

In this technical report, I present a lightweight model for the APSIPA ASC 2025 Challenge, trained using audio-text comparisons. My submission consists of an audio encoder, a text encoder, and an ensemble model. Experiments show that my solution slightly outperforms the baseline on the validation set.

## REFERENCES

[1] J. Bai, M. Wang, H. Liu, *et al.*, *Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift*, 2024. arXiv: 2402.02694.

[2] Y. Wu, K. Chen, T. Zhang, *et al.*, *Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation*, 2024. arXiv: 2211.06687.

[3] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, *Squeeze-and-excitation networks*, 2019. arXiv: 1709.01507.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.