Enhancing Audio Retrieval with Attention-based Encoder for Audio Feature Representation

Feiyang Xiao[1], Qiaoxi Zhu[2], Jian Guan[1*], Wenwu Wang[3]
[1]Group of Intelligent Signal Processing, College of Computer Science and Technology,
Harbin Engineering University, Harbin, 150001, China
[2]Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Ultimo, NSW 2007, Australia
[3]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK

***Abstract*—Pretrained audio neural networks (PANNs) has been successful in a range of machine audition applications. But its limitation in recognising relationships between acoustic scenes and events impacts its performance in language-based audio retrieval, which retrieves audio signals from a dataset based on natural language textual queries. This paper proposes the attention-based audio encoder to exploit contextual associations between acoustic scenes/events, using self-attention or graph attention with different loss functions for language-based audio retrieval. Our experimental results show that the proposed attention-based method outperforms most of state-of-the-art methods, with self-attention performing better than graph attention. In addition, the selection of different loss functions (i.e., NT-Xent loss or supervised contrastive loss) does not have as significant an impact on the results as the selection of the attention strategy.**

***Index Terms*—Language-based audio retrieval, audio representation, attention mechanism, multimodal learning**

I. Introduction

Language-based audio retrieval is a multimodal task that utilizes a text query (i.e., caption) to retrieve the matched audio signal from a provided database [1]–[3]. It benefits search engines to output audio signals matching a text query [3] and enhances the experience of human-machine interaction with improved machine understanding of the audio content [4].

This task is launched by the Detection and Classification of Acoustic Scenes and Events (DCASE) 2022 Challenge Task 6B [2]. The official baseline [2] has the convolutional recurrent neural network (CRNN) [5] as the audio encoder and Word2Vec [6] as the text encoder. The audio encoder extracts audio features (i.e., audio embedding) to represent the acoustic scenes, and the text encoder extracts text features (i.e., sentence embedding) to represent the semantic information of text queries. However, the simple model structure leads the CRNN-based audio encoder's ineffectiveness in audio feature representation and the Word2Vec-based text encoder's ineffectiveness in the semantic information representation, which limits the audio retrieval performance.

The state-of-the-art methods adopt a more complex structure with the pre-trained models for the language-based audio retrieval task [3], [7]–[9]. Specifically, these methods employ pretrained audio neural networks (PANNs) [10] as the audio encoder and BERT-based language model [11] as the text encoder. They achieved improved audio retrieval utilising the pre-trained knowledge from the large-scale audio dataset (i.e., AudioSet [12]) and text datasets (i.e., BooksCorpus [13] and Wikipedia [11]). With the BERT-based language model, the text encoder can capture the contextual information in the text for better semantic information representation. In contrast, the PANNs-based audio encoder often focuses on audio pattern recognization, which has advantages for sound event/scene detection. But PANNs cannot well exploit the contextual association between the acoustic scenes and events within the audio signal due to its convolution operation [14], [15].

Thus, the PANNs-based audio encoder insufficiently represents audio content and limits audio retrieval performance because of mismatching between audio representation and text embedding (involving the contextual information). To address this limitation, our DCASE 2022 Challenge Task 6B submission [16] introduced graph attention network (GAT) [17] in addition to PANNs as the audio encoder to exploit the contextual association of the extracted audio features, and employed Word2Vec [6] as the text encoder. Our submission achieved the 8th place in the DCASE 2022 Challenge Task 6B.

This paper further explores the attention-based audio encoder to capture the contextual association within the audio signal while adopting a BERT-based text encoder to obtain sentence embedding for audio retrieval. Specifically, two different attention mechanisms (graph attention [17] and self-attention [18]) is respectively employed as the attention module, in addition to PANNs in the audio encoder. The graph attention learns the audio feature nodes relation, while the self-attention is widely used for the sequential signal modelling. Experimental result shows that the proposed attention-based method can achieve competitive performance with the state-of-the-art methods, and the ablation studies verify the effectiveness of the attention-based audio encoder. In addition, two different loss functions (NT-Xent loss [19] and supervised contrastive loss [20]) are compared. The result shows that the loss function selection may be less important than the selection of attention mechanisms.

* Corresponding author
This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010, and a GHfund with Grant No. 202302026860.

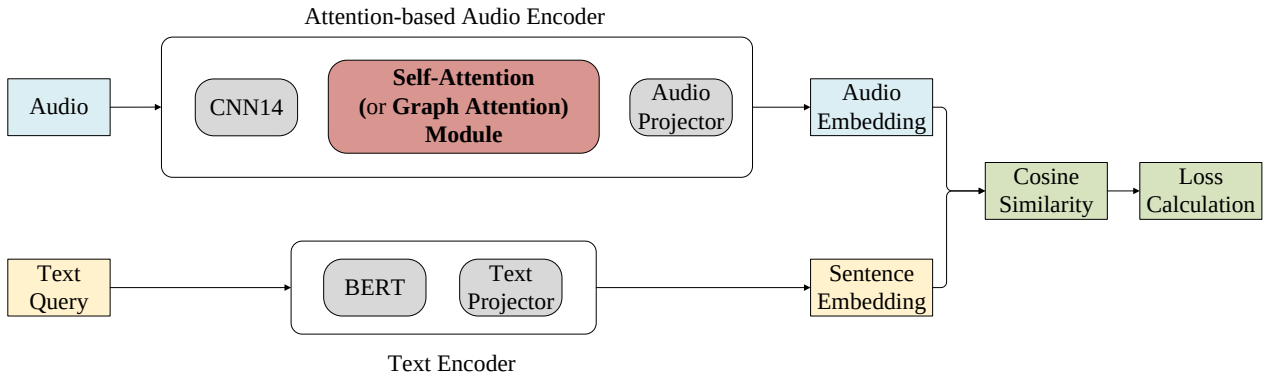

Fig. 1. Framework of the proposed audio retrieval method with the attention-based audio encoder, using either self-attention or graph attention.

II. Proposed Methods

This section introduces the proposed audio retrieval method with the attention-based audio encoder, as illustrated in Figure 1. Specifically, we explore two attention mechanisms, graph attention and self-attention, applied in the audio encoder in addition to PANNs, and eventually formed graph attention based audio retrieval and self-attention based audio retrieval, respectively. In addition, the use of different loss functions is also explored.

A. Attention-based Audio Encoder

We use the attention-based audio encoder to extract audio embedding with semantic information about acoustic scenes/events of the audio signal. The attention-based audio encoder contains PANNs (i.e., CNN14 in [10]) to extract the audio feature, an attention module to capture the contextual association within the audio feature, and an audio projector to output the audio embedding. We consider two attention mechanisms in the attention module, either graph attention or self-attention. The audio feature extracted by PANNs is denoted as $\mathbf{F} \in \mathbb{R}^{T \times D_A}$, where T and D_A denote the time dimension and the latent dimension of the audio feature, respectively.

*1) **Graph Attention Based Audio Encoder**:* The graph attention based audio encoder employs the graph attention network layer as the attention module to capture the relation between audio feature frames (nodes). In the graph attention network layer, the audio feature $\mathbf{F}$ is divided into T audio feature frames, that $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_t, \cdots, \mathbf{f}_T]^\top$ with $1 \leq t \leq T$ and $\top$ denotes the transposition operation. The relation coefficient between two audio feature frames $\mathbf{f}_i$ and $\mathbf{f}_j$ $(1 \leq i, j \leq T)$ is calculated as

$$r_{i,j} = LeakyReLU\left(\mathbf{W}_2 \left[\mathbf{W}_1 \mathbf{f}_i; \mathbf{W}_1 \mathbf{f}_j\right]\right), \tag{1}$$

where matrix $\mathbf{W}_1 \in \mathbb{R}^{D_A \times D_A}$ is learnable to map the audio feature frame with their differences enhanced, operator $[\cdot;\cdot]$ denotes the concatenation of two vectors at the feature vector dimension, and $\mathbf{W}_2 \in \mathbb{R}^{1 \times 2D_A}$ is a learnable vector to map the concatenation result to a relation coefficient. The leaky ReLU function is used for activation. The relation coefficient reflects the degree of the contextual temporal relation between two audio feature frames. All relation coefficients form the relation coefficient matrix $\mathbf{R} \in \mathbb{R}^{T \times T}$. The audio embedding from the graph attention based audio encoder is

$$\mathbf{a} = Proj_A\left\{MeanPooling\left[\mathbf{F} + \sigma(\mathbf{R})\mathbf{F}\mathbf{W}_1^\top\right]\right\}, \tag{2}$$

where $\sigma(\cdot)$ denotes the softmax function, the mean pooling operation is used to squeeze the time dimension, and $Proj_A\{\cdot\}$ is the audio projector using a linear layer to map the squeezed audio feature into the audio embedding $\mathbf{a} \in \mathbb{R}^D$ and D is the dimension of the embedding vector. The residual connection $\left[\mathbf{F} + \sigma(\mathbf{R})\mathbf{F}\mathbf{W}_1^\top\right]$ remains both the information about sound events captured by PANNs and the relation between audio feature frames, thus can learn the contextual association of the audio signal.

*2) **Self-Attention Based Audio Encoder**:* The self-attention based audio encoder employs the self-attention layer [18] as the attention module to capture the temporal relationship within the audio feature. The audio feature $\mathbf{F}$ is mapped into three latent features, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T \times D_A}$ by learnable parameter matrices, $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D_A \times D_A}$, respectively. Then, these latent features are fed to sequential modelling of the self-attention mechanism to learn the temporal relationship, and the residual connection is used to retain the acoustic scenes and events information. So that the audio feature with the temporal information is

$$\hat{\mathbf{F}} = \mathbf{F} + \sigma\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{D_A}\right)\mathbf{V}. \tag{3}$$

The audio embedding from the self-attention based audio encoder can be calculated as

$$\mathbf{a} = Proj_A\left\{MeanPooling\left[\phi(\hat{\mathbf{F}}) + \hat{\mathbf{F}}\right]\right\}, \tag{4}$$

where $\phi(\cdot)$ is a multi-layer perceptron with a depth of two, and the mean pooling operation and the audio projector are used to obtain the audio embedding $\mathbf{a} \in \mathbb{R}^D$. The residual connection remains the learnt temporal relationship and the global information of the audio feature; thus the audio embedding can obtain the contextual association of the audio signal.

B. Text Encoder

Following [3], [7]–[9], we employ the pretrained BERT model [11] as the text encoder of the proposed method to extract the sentence embedding representing the content information of the caption. The caption with L words forms $(L+2)$ tokens, that BERT prepends a "[CLS]" token to obtain the global information of the sentence, and appends a "[SEP]" token to represent the end of the sentence. These tokens are converted as $(L+2)$ token feature vectors with the feature dimension D_S from the pretrained BERT model. Because the "[CLS]" token contains the global information of the sentence, we choose the token feature vector of "[CLS]" to represent the content of the caption sentence. Because the dimension D_S is usually different from the embedding dimension D, a text projector that consists of a linear layer maps the token feature vector of "[CLS]" into the sentence embedding $\mathbf{s} \in \mathbb{R}^D$.

C. Loss Function

The cosine similarity represents the matching degree between audio and sentence embeddings, and the audio signal with a larger cosine similarity matches the caption better. Suppose $\mathcal{B}$ pairs of text query with the corresponding audio signal. There are $\mathcal{B}$ audio embeddings and $\mathcal{B}$ sentence embeddings. The cosine similarity between any sentence embedding $\mathbf{s}_m$ $(1 \leq m \leq \mathcal{B})$ and any audio embedding $\mathbf{a}_n$ $(1 \leq n \leq \mathcal{B})$ is

$$c_{m,n} = \frac{\mathbf{s}_m^\top \mathbf{a}_n}{\|\mathbf{s}_m\|_2 \|\mathbf{a}_n\|_2}, \tag{5}$$

where $\|\cdot\|_2$ denotes the l_2 norm operation. All the $\mathcal{B} \times \mathcal{B}$ cosine similarity values $c_{m,n}$, $1 \leq m \leq \mathcal{B}$ and $1 \leq n \leq \mathcal{B}$, form the cosine similarity matrix $\mathbf{C} \in \mathbb{R}^{\mathcal{B} \times \mathcal{B}}$.

*1) **NT-Xent Loss**:* Following most audio retrieval methods, e.g. [3], [8], [9], [21], we consider the NT-Xent loss [19] to optimize the model during the training stage. It aims to maximise the diagonal elements of $\mathbf{C}$ while minimising the non-diagonal elements of $\mathbf{C}$, that

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{\mathcal{B}} \sum_{m=1}^{\mathcal{B}} \left(\log \frac{e^{\frac{c_{m,m}}{\tau}}}{\sum_{n=1}^{\mathcal{B}} e^{\frac{c_{m,n}}{\tau}}} + \log \frac{e^{\frac{c_{m,m}}{\tau}}}{\sum_{n=1}^{\mathcal{B}} e^{\frac{c_{n,m}}{\tau}}} \right), \tag{6}$$

where τ is the temperature parameter to scale the difference in the cosine similarity value to highlight the difference between different text-audio pairs. The NT-Xent loss assumes that the diagonal elements of $\mathbf{C}$ correspond to text-audio pairs that match, and that the non-diagonal elements correspond to pairs that do not match. However, there may be some non-diagonal elements that actually correspond to matched pairs, but the NT-Xent loss disregards them. Hence, we propose to replace the NT-Xent loss with the supervised contrastive loss [20].

*2) **Supervised Contrastive Loss**:* It marks all the elements of $\mathbf{C}$ corresponding to matched text-audio pairs, and enhances these elements by

$$\mathcal{L}_{\text{supervised}} = -\frac{1}{\mathcal{B}} \sum_{m=1}^{\mathcal{B}} \frac{1}{|K(m)|} \sum_{k \in K(m)} \left(\log \frac{e^{\frac{c_{m,k}}{\tau}}}{\sum_{n=1}^{\mathcal{B}} e^{\frac{c_{m,n}}{\tau}}} + \log \frac{e^{\frac{c_{k,m}}{\tau}}}{\sum_{n=1}^{\mathcal{B}} e^{\frac{c_{n,m}}{\tau}}} \right), \tag{7}$$

where $K(m)$ denotes the set that includes the indices of the matched audio samples of the m-th text query, $|K(m)|$ denotes the number of the indices in $K(m)$, and k is an index from $K(m)$. In this paper, the proposed method using supervised contrastive loss is named with the suffix "+supervised" and the effect of the loss function is discussed in Section III-F.

III. Experimental Results

A. Dataset

We conduct the experiments on the Clotho dataset [22], using both the development split and the validation split as the training set and the evaluation split as the test set. Note that every audio signal has five corresponding captions in the Clotho dataset. The sample rate of the audio signals is 44.1 kHz, and the log-Mel spectrogram of the audio signal is extracted as the input of the audio encoder. We set the dimension of the log-Mel band as 64 and the Hamming window with 50% overlapping while extracting the log-Mel spectrogram.

B. Experimental Setup

The CNN14 module in the audio encoder is initialised by the pretrained parameters from acoustic pattern recognition tasks on AudioSet. The BERT module in the text encoder is initialised by the pretrained parameters from the natural language processing tasks on BooksCorpus and Wikepedia. For the audio encoder, D_A is set as 1024 to fit the output dimension of CNN14. For the text encoder, D_S is set as 768 to fit the output dimension of BERT. The dimension D of the audio and sentence embeddings is 1024. The batch size is set as 60, and the model is optimized by the Adam optimizer [23] with a learning rate of 0.0001. The SpecAugment strategy is used to augment the log-Mel spectrogram for better performance, inspired by [24]. For the NT-Xent loss and supervised contrastive loss, the temperature parameter τ is set as 0.07 following [9].

C. Evaluation Metrics

Following DCASE 2022 Challenge Task 6B [2], we employ the matched audio signal's recall and mean average precision (mAP) metrics to evaluate the retrieval performance. Specifically, the recall at the top-1, top-5 and top-10 retrieved audio signal candidates (R1, R5 and R10, respectively) and the mAP at the top-10 retrieved audio signal candidates (mAP10). Note that the mAP10 metric is the most important metric to rank the submissions of DCASE 2022 Challenge Task 6B.

TABLE I
PERFORMANCE COMPARISON BETWEEN THE STATE-OF-THE-ART METHODS AND THE PROPOSED METHOD (I.E., GRAPH ATTENTION BASED AND SELF-ATTENTION BASED AUDIO RETRIEVAL, THE CNN14 BASED AUDIO RETRIEVAL FOR ABLATION STUDY AND THE SELF-ATTENTION BASED+SUPERVISED FOR THE DISCUSSION ABOUT THE CHOICE OF LOSS FUNCTIONS).

Method	R1(%)	R5(%)	R10(%)	mAP10(%)
DCASE official baseline [2]	2.6	10.2	17.6	6.1
P-GAT [16]	7.0	21.0	33.0	13.0
ATAE-NP-F [3]	7.2	22.5	32.5	13.9
PaSST-MPnet [7]	13.4	35.5	48.2	22.9
RELAX [25]	13.7	35.4	48.4	23.1
SMBO [8]	14.5	37.2	51.0	24.3
Mei_Surrey [9]	14.7	37.7	49.5	24.4
SJTU [21]	**16.2**	38.3	**52.0**	**25.8**
CNN14 based audio retrieval	13.7	35.6	48.7	23.2
Graph attention based	14.6	36.2	50.2	23.9
Self-attention based	14.1	38.0	51.3	24.3
Self-attention based+supervised	14.0	**38.5**	**52.0**	24.4

D. Performance Comparison

We evaluate the retrieval performance by comparing the proposed methods (i.e., graph attention based audio retrieval and self-attention based audio retrieval) with the state-of-the-art methods [3], [7]–[9], [16], [21], [25] and DCASE official baseline [2]. Here, SJTU [21], Mei_Surrey [9], RELAX [25], PaSST-MPnet [7] and ATAE-NP-F [3] respectively achieved the 1st, 2nd, 3rd, 4th and 7th place in DCASE 2022 Challenge Task 6B. P-GAT [16] is our previous submission of Task 6B and achieved 8th place. SMBO [8] is a state-of-the-art method exploring data augmentation's effect on language-based audio retrieval. The performance results are shown in Table I, where "graph attention based" and "self-attention based" are short for graph attention based audio retrieval and self-attention based audio retrieval, respectively.

Table I shows that the proposed graph attention based audio retrieval and self-attention based audio retrieval can achieve better retrieval performance than DCASE baseline, RELAX, ATAE-NP-F, PaSST-MPnet, SMBO and P-GAT methods in terms of all evaluation metrics, except the R1 performance for SMBO. Meanwhile, the proposed methods achieve competitive performance compared with the 1st and 2nd rank methods (SJTU and Mei_Surrey). Moreover, the proposed method has a simpler training processing than the 1st rank method (SJTU) as it does not need pretraining on the AudioCaps dataset [26]. These results show that the proposed graph attention based and self-attention based audio retrieval are effective solutions for language-based audio retrieval.

E. Effect of Attention

To evaluate the effectiveness of the attention-based audio encoder for language-based audio retrieval, we performed an ablation study. Specifically, we removed the attention module (i.e., the graph attention layer in graph attention-based retrieval and the self-attention layer in self-attention-based retrieval) from the proposed encoder, resulting in a degraded audio retrieval method, CNN14 based audio retrieval in Table I. Note that all the settings for the three methods (graph attention-based, self-attention-based and CNN14 based methods) were identical, except for the model structure of the audio encoder.

The ablation study in the second part of Table I shows the graph attention based and the self-attention based audio retrieval outperform the CNN14 based audio retrieval in terms of all evaluation metrics. This indicates that the use of the attention module is the particular element of the proposed methods leading to improved performance, and the contextual association of the audio signal is important to language-based audio retrieval.

In addition, the self-attention based audio retrieval outperforms the graph attention based audio retrieval in terms of all metrics, except the R1 metric. It shows that the self-attention layer is a better choice for the implementation of the attention module in the proposed attention-based audio encoder.

F. Effect of Loss Function

We tested different loss functions (NT-Xent loss and supervised contrastive loss) for the proposed method. Here, we conducted the experiment based on self-attention based audio retrieval and replaced the NT-Xent loss with the supervised contrastive loss, which forms "self-attention based+supervised" in Table I. It has improvements in all metrics except the R1 metric compared with the self-attention based audio retrieval, and has the best performance in terms of R5 and R10 metrics, compared with the state-of-the-art methods. However, the core metric mAP10 is nearly not improved. Therefore, the impact of the loss function is less significant than that of the model structure design, according to those evaluation metrics. A possible reason is that the Clotho dataset ignores that a caption can have multiple matched audio signals, as an audio signal can have multiple corresponding captions, thus some matching text-audio pairs are not marked, thereby affecting the performance.

IV. CONCLUSION

In this work, we propose a language-based audio retrieval method with the attention-based audio encoder, where the attention-based audio encoder is used to capture the contextual association of the audio signal by the attention mechanism. Specifically, we explore two different attention mechanisms, i.e., graph attention and self-attention mechanisms for the attention-based audio encoder, respectively. Experimental results show the proposed method achieves competitive performance with the state-of-the-art methods, verifying the effectiveness of the proposed attention-based audio encoder for audio retrieval and pointing out that the self-attention mechanism is a better choice for the proposed method. Moreover, we discuss the impact of different loss functions, i.e., NT-Xent loss and supervised contrastive loss. Results show that the choice of the attention mechanism has a more significant impact than the selection of loss functions.

REFERENCES

[1] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8867–8871.

[2] H. Xie, S. Lipping, and T. Virtanen, "Language-based audio retrieval task in DCASE 2022 challenge," in *Proc. of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE 2022)*, Nancy, France, November 2022.

[3] B. Weck, M. P. Fern'andez, H. Kirchhoff, and X. Serra, "Matching text and audio embeddings: Exploring transfer-learning strategies for language-based audio retrieval," in *Proc. of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE 2022)*, Nancy, France, November 2022.

[4] F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Local information assisted attention-free decoder for audio captioning," *IEEE Signal Processing Letters*, vol. 29, pp. 1604–1608, 2022.

[5] X. Fu, E. Ch'ng, U. Aickelin, and S. See, "CRNN: A joint neural network for redundancy detection," in *Proc. of International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2017, pp. 1–8.

[6] G. C. Tomas Mikolov, Kai Chen, "Efficient estimation of word representations in vector space," in *Proc. of International Conference on Learning Representations (ICLR)*, 2013.

[7] T. Pellegrini, "Language-based audio retrieval with textual embeddings of tag names," in *Proc. of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE 2022)*, Nancy, France, November 2022.

[8] P. Primus and G. Widmer, "Improving natural-language-based audio retrieval with transfer learning and audio & text augmentations," in *Proc. of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE 2022)*, Nancy, France, November 2022.

[9] X. Mei, X. Liu, H. Liu, J. Sun, M. D. Plumbley, and W. Wang, "Language-based audio retrieval with pre-trained models," DCASE 2022 Challenge, Tech. Rep., July 2022.

[10] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.

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

[13] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 19–27.

[14] F. Xiao, J. Guan, Q. Zhu, and W. Wang, "Graph attention for automated audio captioning," *IEEE Signal Processing Letters*, 2022 (submitted).

[15] H. Song, S. Deng, and J. Han, "Exploring inter-node relations in cnns for environmental sound classification," *IEEE Signal Processing Letters*, vol. 29, pp. 154–158, 2022.

[16] F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Language-based audio retrieval with pretrained CNN and graph attention," DCASE 2022 Challenge, Tech. Rep., July 2022.

[17] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. of International Conference on Learning Representations (ICLR)*, 2018.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, vol. 30. Curran Associates, Inc., 2017.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. of Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

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

[22] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. Zhao, S. Li, T. Ko, H. L. Tang, X. Shao, M. D. Plumbley, and W. Wang, "An encoder-decoder based audio captioning system with transfer and reinforcement learning," in *Proc. of the 6th Detection and Classification of Acoustic Scenes and Events (DCASE 2021) Workshop*, November 2021.

[25] T. L. de Gail and D. Kicinski, "Take it easy: Relaxing contrastive ranking loss with CIDEr," DCASE 2022 Challenge, Tech. Rep., July 2022.

[26] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 119–132.