ENSEMBLE LEARNING FOR AUDIO CAPTIONING WITH GRAPH AUDIO FEATURE REPRESENTATION

Technical Report

Feiyang Xiao[1], *Jian Guan*[1*], *Haiyan Lan*[1], *Qiaoxi Zhu*[2], *and Wenwu Wang*[3]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Centre for Audio, Acoustic and Vibration, University of Technology Sydney, Ultimo, Australia
[3]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

ABSTRACT

This technical report describes our submission for Task 6A of the DCASE2022 Challenge (automated audio captioning). Our system is built based on the ensemble learning strategy. It integrates the advantages of different audio captioning methods, including the graph attention-based audio feature representation method. Experiments show that our ensemble system can achieve the SPIDE_r score (used for ranking) of 30.2(%) on the evaluation split of the Clotho-v2 dataset.

***Index Terms*—** Automated audio captioning, Ensemble learning, Graph attention, Transformer.

1. INTRODUCTION

Automated audio captioning is an intermodal translation task that translates an input audio signal into a corresponding description with natural language (i.e., captions) [1, 2]. AAC is different from the sound event detection (SED) and the acoustic scene classification (ASC) tasks. AAC does not predict a sound event/scene, but describes the general information, including the identification of sound events, acoustic scenes, foreground versus background discrimination, concepts and physical properties of objects and environments [2]. AAC has positive impacts in various applications, such as intelligent and content-oriented man-machine interaction and automatic content description [3], i.e., subtitling television content for the hearing-impaired, analyzing sound scenes for security surveillance.

Existing methods usually employ an encoder-decoder structure, where the audio encoder is used to extract the audio feature, and the decoder generates captions from the audio feature. The initial methods have limited captioning performance due to data insufficiency [4, 5]. With the help of the pretrained audio neural networks (PANNs) [6], the audio encoders of audio captioning methods are effectively improved. The PANNs module is pretrained on a large audio pattern recognition dataset, i.e., AudioSet [7]. Most recent state-of-the-art methods use the PANNs modules as their encoder, and achieve significant improvement in audio captioning task [3, 8, 9, 10, 11]. In addition, to leverage the advantages of different methods, ensemble learning [12] is widely used in the submitted systems for Task 6A of DCASE 2021 Challenge [13, 14].

Corresponding author.

This work was supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010.

Our system includes a method with the encoder-decoder structure using graph attention, namely GraphAC [15]. The audio encoder includes a PANNs module and a graph attention network (GAT) module [16], to model the audio feature with both local and long time dependencies. A Transformer based decoder is used as the decoder in our system. The model is also pretrained on the external dataset, i.e., AudioCaps [17] to further improve the performance. In addition, the ensemble learning strategy [12] is used to improve the final performance of our system.

2. PROPOSED ENSEMBLE SYSTEM

The structure of our ensemble system is shown in Figure 1. It includes five methods to vote the probabilities of the generated words, which can benefit the advantages of these models and improve the captioning performance. Specifically, we incorporate the GraphAC method [15], the GraphAC w/ RL method, the GraphAC w/o top-k method, the P-LocalAFT method [11] and the P-Transformer method [3] to build our ensemble system. The details of these methods and the voting process are introduced as follows.

2.1. GraphAC, GraphAC w/ RL, and GraphAC w/o Top-k

We use an audio feature representation with graph attention learning in the GraphAC, GraphAC with reinforcement learning (i.e., GraphAC w/ RL), and GraphAC w/o top-k methods. The audio encoder uses a graph attention module in addition to the PANNs module (i.e., CNN10). The GraphAC and GraphAC w/ RL methods use the top-k mask to select the audio feature node relations in the adjacency graph, while the GraphAC w/o top-k does not employ this process. In addition, the GraphAC w/ RL method is fine-tuned by reinforcement learning. All the methods with graph attention, i.e., GraphAC, GraphAC w/ RL and GraphAC w/o top-k, use a Transformer based decoder, which has 2 Transformer decoder layers with 8 attention heads.

2.2. P-LocalAFT

The P-LocalAFT [11] employs the PANNs module (i.e., CNN10) as the encoder to extract audio features and designs the LocalAFT decoder to generate captions from audio features. The P-LocalAFT method can incorporate the global and local information of an audio signal in its LocalAFT decoder. The LocalAFT decoder has two layers, and the local region window size is set as 80.

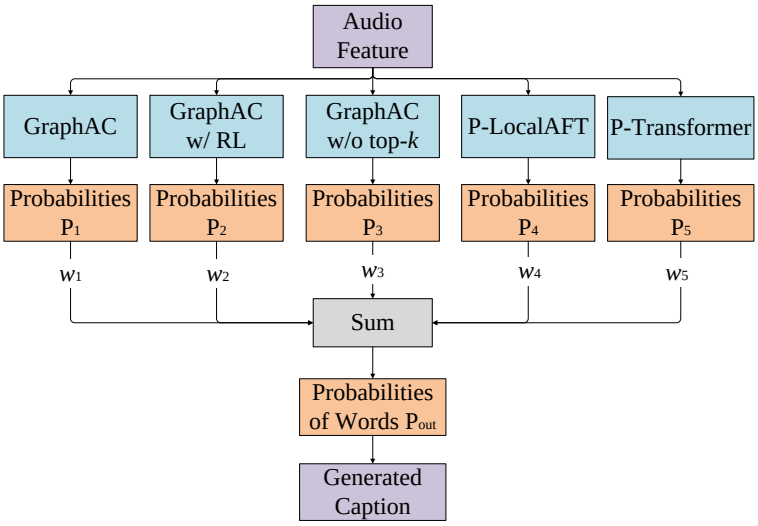

Figure 1: The structure of our proposed ensemble system.

Table 1: Performance comparison on the evaluation split of the Clotho-v2 dataset.

Method	$BLEU_1(\%)$	$BLEU_2(\%)$	$BLEU_3(\%)$	$BLEU_4(\%)$	$ROUGE_l(\%)$	METEOR(%)	$CIDE_r(\%)$	SPICE(%)	$SPIDE_r(\%)$
Baseline system	55.5	35.8	23.9	15.6	36.4	16.4	35.8	10.9	23.3
P-Transformer [3]	56.5	37.2	25.3	17.1	37.7	17.2	41.3	12.3	26.8
P-LocalAFT [11]	57.8	38.7	26.7	17.9	39.0	17.7	43.4	12.2	27.8
GraphAC w/o top-k	58.0	38.8	26.5	17.7	38.4	17.8	43.5	12.4	27.9
GraphAC	58.1	38.6	26.5	18.1	38.5	17.5	43.7	12.6	28.1
GraphAC w/ RL	**66.0**	42.4	27.9	17.0	41.0	17.8	44.2	12.9	28.5
Ensemble system	64.9	**43.9**	**30.3**	**19.9**	**41.5**	**18.1**	**47.1**	**13.3**	**30.2**

2.3. P-Transformer

The P-Transformer is a method in [3], which uses the PANNs module (i.e., CNN10) as the encoder to extract audio features and the Transformer decoder to generate captions from audio features. In our ensemble system, the P-Transformer method does not involve reinforcement learning for fine-tuning. The Transformer decoder has 2 Transformer decoder layers with 8 heads.

2.4. Vote for Results

The audio feature is used as input for the above methods, and then these methods output the predicted probabilities of words, i.e., P_1, P_2, P_3, P_4, and P_5 in Figure 1. In order to leverage the advantages of the above five methods, these probabilities are summarised with their weights (i.e., w_1, w_2, w_3, w_4, and w_5), and then the ensemble system gets the predicted probabilities of words result, i.e., P_{out},

$$P_{out} = w_1 P_1 + w_2 P_2 + w_3 P_3 + w_4 P_4 + w_5 P_5. \quad (1)$$

Finally, with the probability P_{out} for the words to be chosen from the vocabulary, we get the generated caption text. Noting that, the weights are set empirically in our experiments as $w_1 = 33.0\%, w_2 = 21.6\%, w_3 = 17.5\%, w_4 = 15.5\%$, and $w_5 = 12.3\%$.

3. EXPERIMENTS

3.1. Data Processing

We use two datasets in the experiments, i.e., Clotho-v2 [2] and AudioCaps [17]. We firstly pretrain our method on the development and validation splits of the AudioCaps dataset, and then fine-tune the method on the development and validation splits of Clotho-v2. The evaluation split of Clotho-v2 is used for metrics evaluation.

In our experiments, log-Mel spectrograms are used as the input audio features, which are obtained from the raw audio signals with a sample rate of 44.1 kHz. We get 64-dimensional log-Mel spectrogram, using a Hamming window with 50% overlap.

We tokenize the captions of the development set. There are no unknown tokens/words since all the words in the development set appear in the validation, evaluation, and test sets. $<$sos$>$, $<$eos$>$ and $<$pad$>$ are employed to denote the start-of-sequence, the end-of-sequence and sequence padding, respectively. In a batch word vectors input, we pad the word vectors to the max length of this batch with $<$pad$>$.

3.2. Setup

The input word embedding used in our ensemble system is from a word2vec language model [18] that is pretrained on the combination of captions from AudioCaps and Clotho-v2. Following [3], all

five methods applied SpecAugment and mix-up strategies to improve generalisation. The cross-entropy loss with label smoothing [19] was used with Adam optimizer [20] to optimize the network. The batch size was 16, and the learning rate was 0.0001. The early stopping strategy is used according to the SPIDE$_r$ value on the validation set during the training of each method. The model training process stops when the SPIDE$_r$ value is not improved for 10 continuous epochs. Machine translation metrics (i.e., BLEU$_n$, ROUGE$_l$ and METEOR) and captioning metrics (CIDE$_r$, SPICE and SPIDE$_r$) are adopted for performance evaluation.

3.3. Performance Evaluation

In this section, we provide the performance evaluation of our ensemble system in Table 1. In addition, we show the performance of each method used in our ensemble system and the baseline system of Task 6A. Comparing these methods, the GraphAC-based methods (i.e., GraphAC, GraphAC w/ RL and GraphAC w/o top-k) can outperform the compared methods in most evaluation metrics, including SPIDE$_r$, which is the metric used for ranking the submissions to the DCASE Challenge.

The ensemble system achieves the best overall audio captioning performance, as compared with all of the above individual methods, by leveraging the advantages of these methods.

4. CONCLUSION

In this technical report, we have described our ensemble system submitted for Task 6A of the DCASE2022 Challenge. Our ensemble system incorporates five AAC methods to improve captioning performance. Experimental results show the significant performance improvement by our system, which can benefit from the advantages of the individual methods. Especially, our ensemble system can achieve the SPIDE$_r$ score at 30.2%.

5. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, U.S.A., Oct. 2017.

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

[3] X. Mei, Q. Huang, X. Liu, G. Chen, J. Wu, Y. Wu, J. Zhao, S. Li, T. Ko, H. Tang, X. Shao, M. D. Plumbley, and W. Wang, "An encoder-decoder based audio captioning system with transfer and reinforcement learning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 206–210.

[4] E. Çakır, K. Drossos, and T. Virtanen, "Multi-task regularization based on infrequent classes for audio captioning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 6–10.

[5] K. Chen, Y. Wu, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pre-trained CNN," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 21–25.

[6] 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.

[7] J. F. Gemmeke, D. P. 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 *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[8] C. P. Narisetty, T. Hayashi, R. Ishizaki, S. Watanabe, and K. Takeda, "Leveraging state-of-the-art asr techniques to audio captioning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 160–164.

[9] Z. Ye, H. Wang, D. Yang, and Y. Zou, "Improving the performance of automated audio captioning via integrating the acoustic and semantic information," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 40–44.

[10] Q. Han, W. Yuan, D. Liu, X. Li, and Z. Yang, "Automated audio captioning with weakly supervised pre-training and word selection methods," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 6–10.

[11] F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Local information assisted attention-free decoder for audio captioning," *arXiv preprint arXiv:2201.03217*, 2022.

[12] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[13] X. Xu, Z. Xie, M. Wu, and K. Yu, "The SJTU system for DCASE2021 challenge task 6: Audio captioning based on encoder pre-training and reinforcement learning," DCASE2021 Challenge, Tech. Rep., July 2021.

[14] Z. Ye, H. Wang, D. Yang, and Y. Zou, "Improving the performance of automated audio captioning via integrating the acoustic and textual information," DCASE2021 Challenge, Tech. Rep., July 2021.

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

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

[17] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, 2019, pp. 119–132.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of International Conference on Learning Representations (ICLR).*, 2013.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2818–2826.

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