AN ENCODER-DECODER BASED AUDIO CAPTIONING SYSTEM WITH TRANSFER AND REINFORCEMENT LEARNING FOR DCASE CHALLENGE 2021 TASK 6

Technical Report

Xinhao Mei[1], *Qiushi Huang*[1], *Xubo Liu*[1], *Gengyun Chen*[2], *Jingqian Wu*[3*], *Yusong Wu*[3†], *Jinzheng Zhao*[1], *Shengchen Li*[3], *Tom Ko*[4], *H Lilian Tang*[1], *Xi Shao*[2], *Mark D. Plumbley*[1], *Wenwu Wang*[1]

[1] University of Surrey, Guildford, United Kingdom
[2] Nanjing University of Posts and Telecommunications, Nanjing, China
[3] Xi'an Jiaotong-Liverpool University, Suzhou, China
[4] Southern University of Science and Technology, Shenzhen, China

ABSTRACT

Audio captioning aims to use natural language to describe the content of audio data. This technical report presents an automated audio captioning system submitted to Task 6 of the DCASE 2021 challenge. The proposed system is based on an encoder-decoder architecture, consisting of a convolutional neural network (CNN) encoder and a Transformer decoder. We further improve the system with two techniques, namely, pre-training the model via transfer learning techniques, either on upstream audio-related tasks or large in-domain datasets, and incorporating evaluation metrics into the optimization of the model with reinforcement learning techniques, which help address the problem caused by the mismatch between the evaluation metrics and the loss function. The results show that both techniques can further improve the performance of the captioning system. The overall system achieves a SPIDEr score of 0.277 on the Clotho evaluation set, which outperforms the top-ranked system from the DCASE 2020 challenge.

Index Terms— audio captioning, transfer learning, sequence-to-sequence model, reinforcement learning

1. INTRODUCTION

An automated audio captioning (AAC) system describes an audio signal using natural language [1], which is a cross-modal translation task involving the technologies of audio processing and natural language processing. Generating a meaningful description for an audio clip not only requires recognizing audio events but also their properties, activities as well as spatial-temporal relationships between different audio objects [2, 3, 4]. Audio captioning could be useful in several applications, such as subtitling for sound in a television program, assisting the hearing-impaired to understand environmental sounds, and analysing sounds in smart cities for security surveillance.

The encoder-decoder architecture with CNN-Transformer was shown to give excellent performance in the DCASE 2020 challenge [2], and thus is chosen as the baseline system in our work. Audio captioning requires extracting features from the audio modality in the encoder and mapping them into the feature space of the language modality in the decoder, which is a challenging cross-modal translation task. Training an end-to-end audio captioning system from scratch becomes even more difficult, when only a small amount of data is available. Thus, in this work, we investigate how pre-trained models can help address this challenge and improve the performance of an audio captioning system. Another problem in the text generation task is the mismatch between the evaluation metrics and the loss function. The evaluation metrics are discrete and non-differentiable, and thus cannot be optimized directly by back-propagation. To address this problem, we introduce reinforcement learning by incorporating the evaluation metrics into the optimisation of the learning

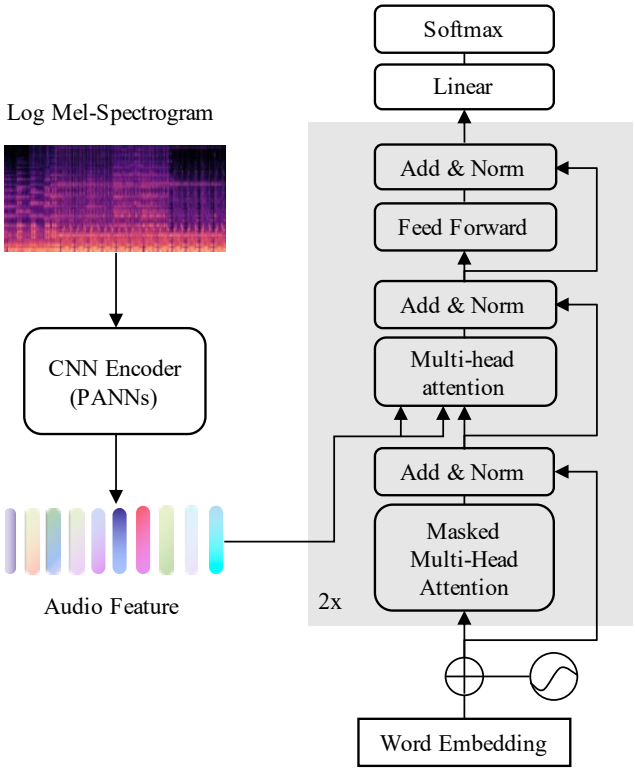

Figure 1: Architecture of the proposed model.

*Jingqian Wu is currently with Wake Forest University, USA
†Yusong Wu is currently with University of Montreal, Canada

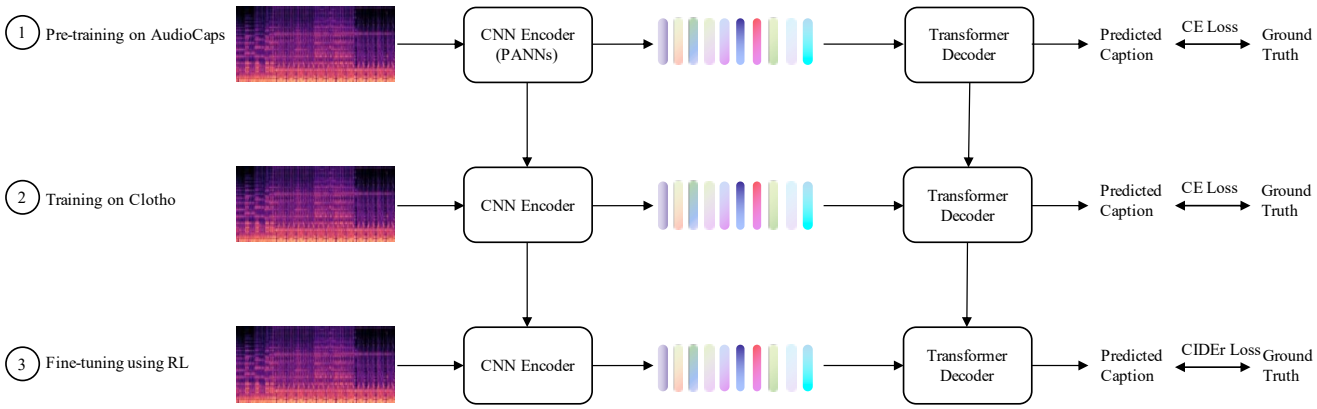

Figure 2: Training stages of the proposed system. The model is firstly pre-trained on AudioCaps dataset, then trained on Clotho dataset. Finally, reinforcement learning is used to optimize the CIDEr score.

system.

This report describes the methods we submitted to Task 6 of DCASE 2021 challenge. Our systems are based on a sequence-to-sequence framework, formed by a CNN encoder and a Transformer decoder, with two additional improvements, including the use of transfer learning from a pre-trained model and the use of reinforcement learning (RL) for optimization on evaluation metrics (e.g. CIDERr).

The remaining sections of this report are organised as follows. In Section 2, the proposed system is described in detail. Experimental setup is presented in Section 3. Results are shown in Section 4. Finally, conclusion are drawn in Section 5.

2. SYSTEM DESCRIPTION

The proposed system is based on a traditional sequence-to-sequence structure which consists of a CNN encoder and a Transformer decoder[1]. This system is then further improved with two techniques. First, transfer learning is introduced to improve the system by using pre-trained models. Second, reinforcement learning is used to optimize the evaluation metric directly. The diagram of the proposed model is shown in Figure 1, while the training procedure is shown in Figure 2.

2.1. CNN encoder

To prevent over-fitting, a relatively simple 10-layer convolutional neural network (CNN) proposed in the pre-trained audio neural networks (PANNs) [5] is used as the encoder of our system. The 10-layer CNN consists of four convolutional blocks where each has two convolutional layers with a kernel size of 3×3. Batch normalization [6] and ReLU nonliearity are used after each convolutional layer. The number of channels in each block are 64, 128, 256 and 512, respectively, and an average pooling layer with kernel size 2×2 is applied between them for down-sampling. Global average pooling is applied along the frequency axis after the last convolutional block and two fully connected layers are followed to further increase the representational ability and to ensure the dimensionality of the output is compatible with the decoder.

[1] https://github.com/XinhaoMei/DCASE2021_task6_v2.git

2.2. Transformer decoder

The decoder is a standard Transformer followed by a classifier [7]. The decoder receives the output of the encoder and outputs a probability distribution along the vocabulary. Transformers are designed to handle sequential data and show stat-of-the-art performance in generation tasks in the area of natural language processing. As the captions in the Clotho dataset are short in length and all of them are between 8 to 20 words, the decoder consists of 2 layers with 4 heads and the dimension of the hidden layer is 128.

2.3. Transfer learning

The use of external data and pre-trained models is allowed in this task, which allows transfer learning to be used. We introduce two transfer learning methods, where the first is transferring from an upstream task while the second is from a larger in-domain dataset.

2.3.1. Pre-trained model for encoder

Different pre-trained neural networks for audio-related tasks have recently been published. PANNs are pre-trained on the AudioSet dataset and show a powerful ability in extracting audio features in different downstream audio pattern recognition tasks [5, 8, 9]. PANNs are used to initialize the parameters of the encoder in all of the experiments.

2.3.2. External data for pre-training

Transfer learning from large datasets to small in-domain datasets is demonstrated to be effective in many tasks. The related work in audio retrieval shows that pre-training the model on the large AudioCaps dataset and then fine-tuning on the Clotho dataset can provide better performance [10]. Inspired by these experimental results, AudioCaps is introduced to pre-train the proposed model and then the model is fine-tuned on the Clotho dataset.

2.4. Fine-tuning via reinforcement learning

The performance of captioning systems can be evaluated by various discrete metrics such as BLUE, CIDEr and SPIDEr [11, 12, 13]. All these metrics are non-differentiable and cannot be directly optimized by back-propagation. Thus, the training objective of the

Model	$BLUE_1$	$BLUE_2$	$BLUE_3$	$BLUE_4$	$ROUGE_L$	METERO	CIDEr	SPICE	SPIDEr
Baseline	0.378	0.119	0.050	0.017	0.263	0.078	0.075	0.028	0.051
PANNs-ZR	0.615	0.403	0.270	0.171	0.392	0.179	0.412	0.122	0.268
PANNs-MR	0.635	0.406	0.268	0.166	0.400	0.176	0.412	0.121	0.266
PANNs-AC-ZW	0.621	0.407	0.273	0.177	0.395	0.179	0.431	0.122	0.277
PANNs-AC-ZR	0.625	0.412	0.278	0.178	0.401	0.176	0.428	0.126	0.277

Table 1: Scores of our submitted models on the Clotho evaluation set. PANNS-ZR and PANNs-MR are all trained on Clotho without using AudioCaps but using randomly-initialized word embeddings, PANNs-ZR uses "zero value masking" as SpecAugment masking type while PANNs-MR uses "mini-batch based mixture masking". PANNs-AC-ZW and PANNs-AC-ZR use AudioCaps to pre-train the model and use "zero value masking" as the masking type of SpecAugment. PANNs-AC-ZW uses the pre-trained word embeddings while PANNs-AC-ZR uses the randomly-initialized word embeddings.

proposed model is to optimize the cross-entropy (CE) loss between the predicted caption and the ground-truth caption. However, the mismatch between the CE loss and the evaluation metrics may lead to performance degradation. To address this issue, reinforcement learning is introduced in our work to optimize the evaluation metric directly, which is then back-propagated in the form of a reward. Previous studies have shown that using rewards from greedy-sampled sentences as the baseline can reduce the high variance of rewards [14]. Subsequent work in [3] showed that the self-critical sequence training (SCST) approach significantly improves the performance on audio captioning tasks, which is used here to optimize CIDEr directly.

3. EXPERIMENTS

3.1. Dataset

3.1.1. Clotho

Clotho is an audio captioning dataset containing a total of 6974 audio samples collected from the Freesound platform and annotated on Amazon Mechanical Turk by annotators from English-speaking countries. To encourage caption diversity, each audio clip is provided with 5 captions annotated by different annotators, thus there are in total 34870 captions. The duration of the audio samples is uniformly ranged from 15 to 30 seconds. Captions are post-processed to make sure there are no unique words, named entities and speech transcriptions.

The test set of Clotho is reserved as the evaluation set for the DCASE challenge. Audio clips in the development set are randomly sampled to create a training set with 5719 audio samples and a validation set with 200 audio samples. During training, each audio clip is combined with one of its five captions as a training sample. During evaluation, all five ground truth captions of an audio clip are used as references and compared with the predicted caption for metric computation.

3.1.2. AudioCaps

AudioCaps is the largest audio captioning dataset created based on AudioSet, which contains around 46k audio samples with duration less than 10 seconds. AudipCaps is divided into three splits, and each audio clip in training set contains one caption, while five captions per audio clip are used in validation and test sets.

3.2. Data pre-processing

64-dimensional log mel-spectrograms using a 1024-points Hanning window with a hop size of 512-points is used as input features. All captions in the dataset are transformed to lower case with punctuation removed. Two special tokens "<sos>" and "<eos>" are used to pad the caption. For the Clotho dataset, the vocabulary contains 4367 words. For transfer learning from AudioCaps to Clotho, these 2 vocabularies are merged together which gives a vocabulary containing 6636 words.

3.3. Experimental setups

The whole model is trained using the Adam optimizer [15] with a batch size of 32. Warm-up is used in the first 5 epochs for the learning rate linearly increased to 0.001. The learning rate is decreased to $1/10$ of itself every 10 epochs after the warm-up. Dropout with rate of 0.2 is applied in the proposed model to mitigate over-fitting problems. To improve the generalization ability of the model, label smoothing is applied in all the experiments [16]. SpecAugment is used with two different making types, "zero-value masking" and "mini-batch based mixture masking" introduced in [17]. Word embeddings are pre-trained by a Word2Vec model using all captions in Clotho and AudioCaps [18].

The model is first trained for 30 epochs, and the model that performs best on the validation set is selected. Then, reinforcement learning is used to optimize the CIDEr score for 25 epochs with a learning rate of 1e-4. During the inference stage, a beam search with a beam size of 3 is used to improve the decoding performance.

4. RESULTS

The challenge allows us to submit up to four different models. Our submission contains the following four models:

- PANNs-ZR: This model is trained only on Clotho using PANNs as encoder. The mask type of SpecAugment is "zero value masking". Word embeddings are random initialized.
- PANNs-MR: This model uses "mini-batch based mixture masking" as make type of SpecAugment and all other settings are same as in the first model.
- PANNs-AC-ZW: This model is first pre-trained on AudioCaps and then fine-tuned on Clotho. The pre-trained word embeddings are used in this model.
- PANNs-AC-ZR: This model is the same as PANNs-AC-ZW except that word embeddings are randomly initialized.

All the models are fine-tuned using reinforcement learning after the training is performed. The performances of our submitted models are shown in Table 1.

5. CONCLUSION

This technical report briefly describes our system and methods for Task 6 of DCASE 2021. Using transfer learning and reinforcement learning, the proposed system significantly improves all evaluation metrics compared to the top-ranked systems in the DCASE challenge last year.

6. ACKNOWLEDGMENT

This work is partly supported by grant EP/T019751/1 from the Engineering and Physical Sciences Research Council (EPSRC), a Newton Institutional Links Award from the British Council, titled "Automated Captioning of Image and Audio for Visually and Hearing Impaired" (Grant number 623805725) and a Research Scholarship from the China Scholarship Council (CSC) No. 202006470010.

7. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2017, pp. 374–378.

[2] K. Chen, Y. Wu, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, "Audio captioning based on transformer and pretrained cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events. Tokyo, Japan*, 2020, pp. 21–25.

[3] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A crnn-gru based reinforcement learning approach to audio captioning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, 2020, pp. 225–229.

[4] K. Nguyen, K. Drossos, and T. Virtanen, "Temporal subsampling of audio feature sequences for automated audio captioning," *arXiv preprint arXiv:2007.02676*, 2020.

[5] 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.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[8] A. Ö. Eren and M. Sert, "Audio captioning based on combined audio and semantic embeddings," in *2020 IEEE International Symposium on Multimedia*. IEEE, 2020, pp. 41–48.

[9] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 101–105.

[10] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," *arXiv preprint arXiv:2105.02192*, 2021.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[12] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[13] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international conference on Computer Vision*, 2017, pp. 873–881.

[14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

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

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

[17] H. Wang, Y. Zou, and W. Wang, "Specaugment++: A hidden space data augmentation method for acoustic scene classification," *arXiv preprint arXiv:2103.16858*, 2021.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.