AUDIO CAPTIONING WITH MESHED-MEMORY TRANSFORMER

Technical Report

Zhiwen Chen[1]*, Dawei Zhang, Jun Wang, Feng Deng*

[1] University of Chinese Academy of Sciences. Beijing, China
chenzhiwen20s@ict.ac.cn

ABSTRACT

Automated audio captioning is the task of describing the audio content of a given audio signal in natural language. Through advancing for years, transformer-based language models are widely used in audio captioning. However, most architectures based on transformer, cannot learn prior knowledge between samples well, leading to worse text decoding. For better acoustic event and language modeling, a sequence-to-sequence model is proposed which consists of a CNN-based encoder, a memory-augmented refiner and a meshed decoder. The proposed architecture refines a multi-level representation of the relationships between audio features integrating learned a priori knowledge. At decoding stage, it exploit low- and high-level features with a mesh-like connectivity. Experiments show that the proposed model can achieve a SPIDEr score of 0.2645 on Clotho V2 dataset.

Index Terms— DCASE 2021, audio captioning, transformer

1. INTRODUCTION

Automated audio captioning is an inter-modal translation task, where a system generates the textual description for a given audio signal [1].

Similar to image captioning, audio captioning is mostly based on an encoder-decoder architecture. Most captioning techniques have employed RNNs as language models and used the output of one or more layers of a CNN to condition language generation [2] . In order to capture visual features better, the feature extractor is usually pre-trained on large-scale classification datasets, like ImageNet. Inspired by that, the PANNs [3] are utilized for audio embeddings, which are pre-trained on the AudioSet [4].

As for text decoding stage, mainstream language model schemes still take transformers as their primary frameworks at present [5]. Transformers are created to excavate long-range dependencies among word embeddings with self-attention operator in thefield of Natural Language Processing(NLP), and achieve great results in sequence generation tasks. However, the basic self-attention has a significant limitation. Because everything depends solely on pairwise similarities, self-attention cannot model a priori knowledge on relationships between samples. To achieve better feature extraction and language modeling, a memory-augmented refiner and a meshed decoder are applied [6]. These models incorporate a region encoding approach that exploits a priori knowledge through memory vectors and a meshed connectivity between encoding and decoding modules.

Experiment results show that the proposed method outperforms the previous baseline model and reached a SPIDEr score of 0.2645 on Clotho V2 dataset for audio captioning task.

2. METHOD

The proposed audio captioning model architecture is based on transformer structure. The modifications are listed as below.

2.1. CNN-based Extractor

In audio captioning challenge, a system is required to understand and model the relationships between audio and textual elements, and then to generate a sequence of corresponding words. This has usually been tackled via better semantic feature capture. Similar to audio captioning, most captioning techniques have employed CNN-based framework to encode visual information in computer vision tasks. Inspired by that, for the input spectrogram, the CNN-based network is adapted and used for audio feature extraction.

The architecture of convolutional neural networks (CNN) has shown great performance on audio pattern recognition. CNN-based methods have achieved state-of-the-art performance in several DCASE challenge tasks, such as acoustic scene classification and sound event detection. The CNN system consists of several convolutional blocks. Each convolutional block contains several kernels that are convolved with the input feature maps to capture their local patterns.

Despite their wide adoption, CNN-based feature extractors suffer from excessive extraction ability and limited receptive field. CNNs with large receptive field degrade in performance and fail to generalize in acoustic scene classification. Considering this factor, we employ a relatively simple CNN, such as a 10-layer CNN, rather than a deeper model [7] .

2.2. Memory-Augmented Refiner

Similar to image captioning, instead of directly feeding CNN features to the decoder, a refining module is proposed, which contains a memory-augmented transformer encoder to refine their representations. In our framework, attention mechanism is utilized to incorporate spatial knowledge on the audio encoding stage. The memory-augmented refiner extends “keys” and “values” set with additional “slots”, which can encode a priori knowledge on relationships between audio features. Unlike traditional self-attention operator, these additional “slots” are designed to capture auxiliary information for inference.

The most basic computation in the transformer, scaled dot-product attention, is adapted for additional memory “slots” [6],

Table 1: Experimental results on Clotho V1 evaluation data

$model$	$Bleu_1$	$Bleu_2$	$Bleu_3$	$Bleu_4$	$METEOR$	$ROUGE_L$	$CIDEr$	$SPICE$	$SPICEr$
Cnn10-Transformer	0.5275	0.3393	0.2213	0.1363	0.1593	0.3623	0.3325	0.1131	0.2228
Cnn14-Transformer	0.5369	0.3493	0.2213	0.1440	0.1602	0.3676	0.3487	0.1119	0.2303
ResNet22-Transformer	0.5409	0.3630	0.2460	0.1578	0.1633	0.3743	0.3689	0.1170	0.2429
Cnn10-M2Transformer	0.5565	0.3635	0.2419	0.1562	0.1667	0.3728	0.3958	0.1129	0.2543
Cnn14-M2Transformer	0.5562	0.3565	0.2359	0.1526	0.1661	0.3594	0.3983	0.1175	**0.2579**
ResNet22-M2Transformer	0.5586	0.3597	0.2374	0.1537	0.1681	0.3655	0.3865	0.1202	0.2533

Table 2: Experimental results on Clotho V2 evaluation data

$model$	$Bleu_1$	$Bleu_2$	$Bleu_3$	$Bleu_4$	$METEOR$	$ROUGE_L$	$CIDEr$	$SPICE$	$SPICEr$
Cnn10-M2Transformer	0.5608	0.3691	0.2493	0.1666	0.1693	0.3727	0.4064	0.1184	0.2624
Cnn14-M2Transformer	0.5550	0.3574	0.2361	0.1525	0.1683	0.3658	0.4088	0.1203	**0.2645**
ResNet22-M2Transformer	0.5671	0.3735	0.2517	0.1647	0.1704	0.3770	0.3984	0.1160	0.2572
ResNet38-M2Transformer	0.5631	0.3671	0.2441	0.1578	0.1698	0.3707	0.4059	0.1190	0.2624

which can be defined as:

$$MemoryAttention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax(\frac{(W_q\boldsymbol{Q})[W_k\boldsymbol{K}, M_k]^T}{\sqrt{d_k}})[W_v\boldsymbol{V}, M_v] \quad (1)$$

where $\boldsymbol{Q} \in R^{n_q \times d_{model}}$, $\boldsymbol{K} \in R^{n_k \times d_{model}}$, $\boldsymbol{V} \in R^{n_v \times d_{model}}$, $\boldsymbol{M_k} \in R^{m \times d_k}$, $\boldsymbol{M_v} \in R^{m \times d_v}$, are the queries, keys, values, slots on keys and slots on values, respectively. $W_q \in R^{d_{model} \times d_q}$, $W_k \in R^{d_{model} \times d_k}$, $W_v \in R^{d_{model} \times d_v}$ are linear layer weights for queries, keys, and values vectors. $[\cdot, \cdot]$ indicates concatenation.

2.3. Meshed Decoder

The decoder used in the proposed model is a meshed transformer decoder, which accepts (refined) feature vectors, then generates a sequence of caption. Different from general decoder, meshed decoder utilizes multi-layer representation of refined features while still constructing the multi-layer structure [6]. The meshed decoder can obtain multi-layer features simultaneously during decoding, which can take advantage of all encoding layers during the generation of the sentence.

In order to capture all encoding layers, meshed cross-attention operator is proposed. We first use keys and values from the encoder to calculate cross-attention along with queries from the decoder, as follows:

$$\Gamma(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y}) = Attention(W_q\boldsymbol{Y}, W_k\tilde{\boldsymbol{X}}^i, W_v\tilde{\boldsymbol{X}}^i) \quad (2)$$

where $\tilde{\boldsymbol{X}}^i$ stands for output from encoding layer i, while $\boldsymbol{Y}$ is the input sequence vector.

To adjust importance of each encoder layer feature dynamically, a matrix of weights α_i is calculated as:

$$\alpha_i = \sigma(W_i[(Y), \Gamma(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y})] + b_i) \quad (3)$$

where $[\cdot, \cdot]$ indicates concatenation, σ represents sigmoid activation, W_i and b_i are parameters of a linear layer.

At last, meshed cross-attention operator can be modified as

$$MeshedAttention(\tilde{\boldsymbol{X}}, \boldsymbol{Y}) = \sum_{i=1}^{N} \alpha_i \odot \Gamma(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y}) \quad (4)$$

3. EXPERIMENTS

3.1. Datasets

We evaluated our model on the Clotho dataset [8], which consists of audio clips from the Freesound platform and its captions were annotated via crowdsourcing. Clotho dataset is an audio captioning dataset, now reached version 2. And all audio samples in the Clotho dataset are of 15 to 30s duration and captions are eight to 20 words long.

For version 2, We used the development and validation split of total 4884 audio clips with 24420 captions (i.e. one audio clip has five ground-truth captions) for training and the evaluation split of 1045 audio clips with 5225 captions for testing. As for version 1, the development split of 2893 audio clips is used for training and the evaluation split of 1045 audio clips is used for testing.

Among the metrics used, BLEUn [9] measures a modified n-gram precision. ROUGEL [10] measures a score based on the longest common subsequence. METEOR [11] measures a harmonic mean of weighted unigram precision and recall. CIDEr [12] measures a weighted cosine similarity of n-grams. SPICE [13] measures the F-score of semantic propositions extracted from caption and reference. SPIDEr [14] is the arithmetic mean between the SPICE score and the CIDEr score. In all metrics used, higher scores indicate better performance.

3.2. Results

Table 1 shows the evaluation results on the Clotho V1 dataset. As it can be seen, M2Transformer method surpasses the current state of the art on all metrics. With respect to Cnn14-Tranformer system, Cnn14-M2Tranformer system achieves an advancement of 2.7 SPICEr points.

The performance of the proposed model on Clotho V2 dataset is shown in Table 2. It can be seen that the proposed model reached a SPIDEr score of 0.2645.

4. CONCLUSION

We present and use M2 Transformer, a novel Transformer-based architecture for audio captioning. Our system learns a multi-

level representation of the relationships between audio features integrating learned a priori knowledge, and uses a mesh-like connectivity at decoding stage to exploit low- and high-level features. Experimental results validate the effectiveness of our proposed approach for audio captioning task.

5. REFERENCES

[1] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2017.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, and M. 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.

[4] J. F. Gemmeke, D. Ellis, D. Freedman, A. Jansen, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics*, 2017.

[5] A. Shi, "Audio captioning with the transformer," DCASE2020 Challenge, Tech. Rep., June 2020.

[6] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] K. Chen, Y. Wu, Z. Wang, X. Zhang, and X. Shao, "Audio captioning based on transformer and pre-trained cnn," in *Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020.

[8] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," 2019.

[9] S. Papineni, "Bleu: A method for automatic evaluation of machine translation," in *Meeting of the Association for Computational Linguistics*, 2002.

[10] C. Y. Lin, "Automatic evaluation of summaries," *US*, 2010.

[11] B. Satanjeev, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," *ACL-2005*, pp. 228–231, 2005.

[12] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *IEEE*, 2015.

[13] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," *Adaptive Behavior*, vol. 11, no. 4, pp. 382–398, 2016.

[14] S. Liu, Z. Zhu, Y. Ning, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.