CAU SUBMISSION TO DCASE 2023 TASK6A: AUDIO CAPTIONING USING WAVEGRAMS THAT CONTAIN FREQUENCY INFORMATION

Technical Report

Seungmin Chou[1], Jaeseung Yim[1], Changwon Lim[1]

[1]Chung-Ang University, Department of Applied Statistics, Seoul, South Korea,
csmwstat@gmail.com, yimyh1231@gmail.com, clim@cau.ac.kr

ABSTRACT

This technical report describes an Automated Audio Captioning model for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge, Task 6A. Utilizing wave gram and patchout as proposed in [1] and [2], respectively, we propose audio captioning using Wavegrams that contain frequency information. We use pre-trained models trained using AudioSet data, to make word embedding. Our proposed sequence-to-sequence model consists of CNN14 encoder and a Transformer decoder. Experiments show that the proposed model achieves a SPIDEr score of 0.011.

Index Terms— Automated Audio Captioning, Wavegram, pretraining, PASST

1. INTRODUCTION

Automated Audio Captioning (AAC) refers to the task that generating captions or transcripts for audio files. In recent years, there has been a growing interest in Audio Automated Captioning. Thanks to the AAC Challenge organized by DCASE, numerous methodologies have been proposed recently [3, 4], and we have paid particular attention to models using patchout.

One of the primary obstacles in AAC is the insufficient amount of data available. To overcome this issue, several modern techniques employ pre-trained models such as PANNs and ResNet, resulting in significant enhancements in the overall performance of the system. To address this challenge, Mei et al. [5] utilized a transformer encoder that was pre-trained on the audio tagging task. Similarly, Kouzelis et al. [6] employed patchout faSt Spectrogram Transformer (PASST) pre-trained on AudioSet data to overcome the lack of training data for AAC. Building on the pre-trained PASST-based transformer proposed by Kouzelis et al., we introduce an AAC model that utilizes Wavegram [1].

The remaining of this report is organized as follows. Section 2 describes our framework and architecture. The experimental setup is given in Section 3. Section 4 presents the results on the evaluation set. Finally, conclusion is drawn in Section 5.

2. SYSTEM DESCRIPTION

Our proposed model's core architecture follows a conventional sequence-to-sequence structure, comprising two feature extractors, a CNN14 encoder and a Transformer decoder. There are two feature extractors in total. One of them is the system of one-dimensional CNNs proposed by Kong et al. (2020) [1] for creating wavegrams, which can learn frequency information that is not present in log mel spectrograms. The other one is a PASST model pre-trained on Audio Set data, which extracts the textual input. The wavegrams are passed as inputs to the encoder, along with the log mel spectrograms that are created from the audio file. The textual input is used as input embeddings for the decoder. The encoder generates an abstract embedding sequence from the input, which is then passed to the decoder to produce an audio caption.

2.1. Wavegram

Wavegram, proposed by Kong et al. (2020) [1] is similar to log Mel spectrogram but learned using a neural network. Wavegram is designed to learn a time-frequency representation that is a modification of the Fourier transform and has a time and frequency axis. It can learn frequency information that may be lacking in one-dimensional CNN systems and may improve over hand-crafted log Mel spectrograms by learning a new kind of time-frequency transform from data. To build a Wavegram, the authors first apply a one-dimensional CNN to the time-domain waveform. The CNN begins with a convolutional layer with a filter length of 11 and stride 5 to reduce the size of the input. This is followed by three convolutional blocks, each consisting of two convolutional layers with dilations of 1 and 2, respectively, which are designed to increase the receptive field of the convolutional layers. Each convolutional block is followed by a downsampling layer with stride 4. By using the stride and downsampling three times, a 32 kHz audio recording is downsampled to 100 frames of features per second.

2.2. Pretrained PASST on AudioSet

Patch out faSt Spectrogram Transformer (PASST), as the name suggests, is a model that applies patching out to spectrogram transformers, proposed by Koutini et al. (2021) [2]. In addition to patch-out, PaSST also uses distinct embeddings for time and frequency positional encoding. This approach offers the benefit of separating time and frequency embeddings, which enables

handling inputs of varying lengths without the need for fine-tuning or interpolation. PaSST has achieved state of the art performance in audio classification tasks. We used a PASST model pre-trained on the Audio Set data to create a logit vector using the 527 classes derived from the Audio Set data and used the logit as word embeddings for the transformer.

2.3. Encoder and decoder

Our model utilizes the CNN14 as an encoder and the transformer as a decoder, both of which were part of the baseline architecture. In addition to log mel spectrograms, we incorporated Wavegrams (concatenated with log mel spectrograms) as inputs to the encoder, following the approach proposed in the PANNS model. Furthermore, we utilized logits extracted from a pretrained PASST model on Audioset data as input embeddings for the decoder.

3. EXPERIMENTAL SETUP

We used the Clotho dataset v2.1 as our main dataset for training and evaluation. The learning rate for our model was set to 1.0e-05, and the training process was conducted over 20 epochs. ADAMW was used as the optimizer for the training process, and cross-entropy was chosen as the loss function. We selected the best model based on the lowest loss value on the validation dataset.

4. RESULT

The experimental results show that the model with a CNN14 encoder and Transformer decoder achieved a SPIDER score of 0.036 and SPIDEr-fl score of 0.011.

5. CONCLUSION

Through this report, we present the results of our submission for Task 6A of the DCASE 2023 Challenge. We propose a model that incorporates not only log Mel spectrograms, but also Wavegram and textual context. Our best model achieved an SPIDEr score of 0.011 on the evaluation dataset.

6. REFERENCES

[1] 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 Transactionson Audio, Speech and Language Processing, vol. 28, pp. 2880–2894, 2020.

[2] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, “Efficient training of audio transformers with patchout,” CoRR, vol. abs/2110.05069, 2021. [Online]. Available: https://arxiv.org/abs/2110.05069

[3] X. Xu, Z. Xie, M. Wu, and K. Yu, “The SJTU system for DCASE2022 challenge task 6: Audio captioning based on encoder pre-training and reinforcement learning,” DCASE2022 Challenge, Technical Report

[4] Z. Ye, Y. Zou, F. Cui, and Y. Wang, “Automated Audio-Captioning with multi-task learning,” DCASE2022 Challenge, Technical Report

[5] X. Mei, X. Liu, H. Liu, J. Sun, M. D. Plumbley, and W. Wang, “Automated Audio Captioning with keywords guidance,” DCASE2022 Challenge, Technical Repor

[6] T. Kouzelis, G. Bastas, A. Katsamanis, A. Potamianos, “Efficient Audio Captioning Transformer with patchout and text guidance”, DCASE2022 Challenge, Technical Report

[7] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7008–7024.