AUDIO CAPTIONING WITH THE TRANSFORMER

Technical Report

Anna Shi

ShuangFeng First
shi.anyan@protonmail.com

ABSTRACT

In this technical report, we present the techniques and models applied to our submission for DCASE 2020 task 6: automated audio captioning. We aim to focus primarily on how to apply transformer methods efficiently to deal with large amount of audio data. Our experiments with the public DCASE2020 challenge task 6 Clotho evaluation data resulted in a SPIDEr of 0.1171, while the SPIDEr of the official baseline is 0.054.

***Index Terms*—** DCASE 2020, audio captioning, transformer

1. INTRODUCTION

The Detection and Classification of Acoustic Scenes and Events (DCASE) is a series of challenges aimed at developing sound classification and detection systems [1]. In this year, task 6 is automated audio captioning, the aim is the task of general audio content description using free text. For the detailed information about the dataset and the challenge, please refer [1].

2. PROPOSED METHODS

2.1. Feature extraction

The dataset for task 6 is Clotho dataset [2, 3]. No preprocessing step was applied in the presented frameworks. The acoustic features for the 44.1kHz original data used in this system consist of 128-dimensional log mel-band energy extracted in Hanning windows of size 2048 with 431 points overlap.

In order to prevent the system from overfitting on the small amount of development data, we added random white noise (before log operation) to the Mel spectrogram in each mini-batch during training. We also propose to introduce data augmentation by temporal-frequency shift. The temporal shift augmentation is a random shift of the signal by rolling the signal along the time axis. The frequency shift augmentation is a random roll in the range +-5% around the frequency axis in the Mel domain. A wrap-around both temporal-frequency shifts to preserve all information. Here +-5% wrap-around in frequency does not affect the sound much physically or perceptually, but can generate a lot of augmented data. One thing to note is that the frame-level labels of strongly annotated synthetic data also have to be shifted accordingly over the temporal shift.

2.2. Model architecture

The transformer [4] is employed for this task. The transformer is an encoder-decoder type of architecture which uses two different attention layers: encoder-decoder attention and self-attention. The input dimension of encoder is 128. The numbers of encoder and decoder stacks are both 6. The number of heads in multi-head attention is 8. The dimensions of key and value are both 64.

2.3. Submissions

For this challenge, We submitted 4 prediction results with different trained models.

- Shi_SFF_task6_1.output.csv: The SPIDEr on evaluation data was 0.106.
- Shi_SFF_task4_2.output.csv: The SPIDEr on evaluation data was 0.108.
- Shi_SFF_task4_3.output.csv: The SPIDEr on evaluation data was 0.117.
- Shi_SFF_task4_4.output.csv: The SPIDEr on evaluation data was 0.116.

3. REFERENCES

[1] http://dcase.community/challenge2020/.

[2] S. Lipping, K. Drossos, and T. Virtanen, "Crowdsourcing a dataset of audio captions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1907.09238

[3] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020. [Online]. Available: https://arxiv.org/abs/1910.09387

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.