SOUND EVENT DETECTION USING CONVOLUTION ATTENTION MODULE FOR DCASE 2023 CHALLENGE TASK4A

Technical Report

Sumi Lee[1]*, Narin Kim*[1]*, Juhyun Lee*[1]*, Chaewon Hwang*[1]*, Sojung Jang*[1]*, Il-Youp Kwak*[1]

[1] Chung-Ang University, Department of Applied Statistics,
Seoul, South Korea, {dltnal821, nrgolden, juhyun0917, ladyikol, thwjd990519, ikwak2}@cau.ac.kr

ABSTRACT

In this technical report, we propose sound event detection models based on CRNN for DCASE 2023 challenge task4A. DCASE task4 evaluates the model with two main metrics. The two metrics are PSDS1 and PSDS2, which have different characteristics, making it difficult to dramatically raise two metrics with one model. Therefore, we have developed two models with different directions. The first model is the Flcam-CRNN, which aimed at PSDS1. Flcam is an attention module created by reflecting the features of 2D audio features in the time-frequency domain. The second model is Mha-CRNN, which aimed at PSDS2. SED data has the characteristic of containing several sounds about a space. Therefore, multi-head attention was used to extract features from various perspectives.

Index Terms— CRNN, attention, convolution module, multi-head attention

1. INTRODUCTION

The objective of DCASE task4A is to detect sound events in each audio clip and predict the onset and offset (localization) of the event. In this technical report, we propose a sound event detection model using CRNN for the DCASE 2023 challenge task4A. The task evaluates the model with two main matrices, psds1 and psds2, which have distinct characteristics. It is not easy to dramatically improve both matrices with a single model. Therefore, we developed two models that aim at psds1 and psds2 respectively.

The first model is flcam-CRNN. Since we need to predict the onset and offset of an event, our goal is to use an attention module to learn the important timestamp of the event. If it were image data, we would use attention modules such as BAM and CBAM to focus on the important parts of an image. However, unlike images, which are shift-invariant along all axes, audio is not shift-invariant along the frequency axis. FDY-CRNN [1] corroborates that applying the same attention module to audio data does not help enhance the performance. Inspired by the idea, we developed Flcam, an audio-specific attentional module. This model improves the performance of psds1 and causes low computational cost through a simple module structure.

The second model is mha-CRNN. Sound Event Detection (SED) data includes multiple sounds occurring simultaneously in a space or sounds with different frequencies at the same time. Considering this aspect, it is effective to extract multiple features of audio data. Therefore, we propose a CRNN model using a multi-head attention module to extract features from different perspectives. This model is appropriate for improving the performance of psds2.

2. METHOD

2.1. Flcam-CRNN

Flcam is a convolutional attention module that reflects the characteristics of 2D audio features in the time-frequency domain. While reflecting the frequency shift-variant features of audio data, we have considered how to assign feature-adaptive attention weight. We converted the wav file into 2D audio features and applied the Conv2D, batch normalization, and context gating layers before applying the flcam attention module. The best performance was obtained when the block with the attention module was repeated 6 times. We thus stacked 6 blocks. Furthermore, the performance gets better when we increase the channel dimension. Therefore, we doubled the channel of the CRNN to improve the representation and further improved the performance of the sound event detection system. After the blocks, it goes through bi-GRU (bidirectional GRU) and then a classifier that outputs a strong label for time and a weak label for class.

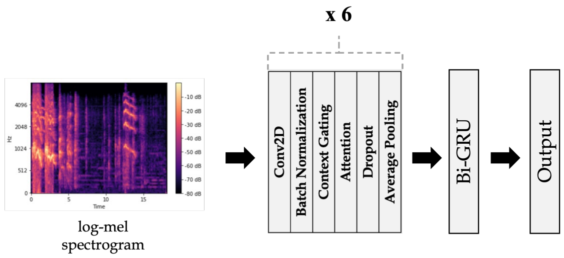

Figure 1: Flcam-CRNN

2.2. Mha-CRNN

As mentioned earlier, SED data consists of several sounds with different frequencies that occur simultaneously. We thought that there would be a limit to capturing all the characteristics of SED data with only one attention head. Mha-CRNN takes it into account and adopts multiple attention heads. They are intended to capture various frequencies that are generated at the same time. After going through conv2D for 2D audio features, a multi-head attention module was applied. Like Flcam-CRNN, Mha-CRNN doubled the channels of CRNN to improve the representation ability.

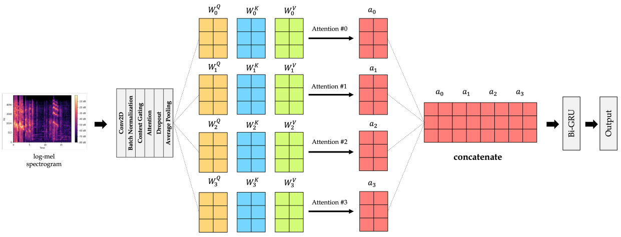

Figure 2: Mha-CRNN

3. EXPERIMENTS

3.1. Data

The audio dataset is primarily based on the DESED dataset, which is a dataset designed to recognize sound event classes in domestic environments. DESED is composed of 10 event classes to recognize in 10 second audio files recorded in domestic environments. The training set consists of (1) synthetic strongly labeled data (10000), (2) unlabeled in domain data (14412)and (3) weakly labeled (1578). The subset of weak data was used for training and remainder is used for validation. Strong real data is external data, and there are two systems which include it to train the network and other except for it. The validation set is (1) synthetic strongly labeled (1168), and the performance of the model was evaluated by strongly labeled.

- train set
  1. synthetic strongly labeled data (10000)
  2. unlabeled in domain data (14412)
  3. weakly labeled (1578)
- validation set
  1. strongly labeled (1168)
- test set
  1. strongly labeled (1168)

3.2. Feature extraction

The audio feature was used by extracting the wav-type audio as a logmel spectrogram after short tiem Fourier transform (stft) with 2048 window length and 256 hop length such as baselines. Therefore, the final input is a logmel spectrogram sampled at 128 dimensions of the mel bin size and a sample rate of 16000. The audio length was set to 10 seconds by padding or cutting.

3.3. Augmentation

Data augmentation techniques transform the data so that the model can perform well in the face of unseen data. In other words, it is a way to increase the generalization power of a model. Data augmentation techniques used in image data include flipping, rotating, cropping, and color jittering [2]. For audio data, you need to use different methods depending on the characteristics of the audio data. The methods we used are mixup, time mask, and filter augmentation.

First one is mixup [3]. Mixup, a data-agnostic data augmentation, extends the training distribution under the assumption that linear interpolations of feature vectors should lead to linear interpolations of the associated targets. This simple method improves generalization of the model. We used soft mixup. In soft mixup, the labels are represented as probability distributions over the possible classes, rather than binary vectors. This allows for a more continuous interpolation between examples and their labels, which can improve the generalization performance of the neural network. The formula for the mixup is following:

$$\tilde{X} = \lambda X i + (1 - \lambda) X j. \tag{1}$$

$$\tilde{Y} = \lambda Y i + (1 - \lambda) Y j. \tag{2}$$

Lambda is a parameter that determines the proportion of data to which the mixup is applied, and was randomly drawn from the beta distribution for each epoch. We chose the hyperparameter of beta distribution, which are alpha and beta both to be 0.2 as they are the most commonly used values when applying mixups in the SED field.

In addition, we used time masking suggested in SpecAugment [4]. This technique masks a portion of the audio feature of log mel spectrogram along the time axis. We randomized the values for where the time mask was applied to give a uniform width of masking.

Lastly, filter augmentation [5], the technique inspired by frequency masking, augments audio data along the frequency axis. Unlike frequency masking, it doesn't completely zero out a specific region of the frequency, but instead creates bands of frequencies and gives them different weights. It processes data to better learn the patterns in 2D audio features by making important information more salient and unnecessary frequencies weaker. To make Mean-Teacher model [6] more robust, this augmentation was applied differently to the student and teacher model to introduce noise into the two models. Filter augmentation was applied once for the input of student model and twice for the input of teacher model.

3.4. Mean Teacher

Since this task has data without timestamps and class labels, we need to use semi-supervised learning scheme to utilize all the data. Our team adopted the Mean Teacher Method [7] as in the given baseline model and applied different augmentations to the student and teacher model.

3.5. Training

The hardware used for training was 1 Quadro RTX 8000. We used Adam optimizer. All models were experimented with up to 200 epochs and early stopping was applied based on the event-based F1 score.

4. RESULTS

We submitted two SED systems to this competition. According to Table 1, Flcam-CRNN has an average psds1 performance of 0.437, which is 0.08 higher than that of the baseline. psds2 also has an average of 0.651 and is improved by 0.09 over the baseline. In case of Mha-CRNN, psds1 is 0.069, which is lower than the baseline. However, psds2 is about 0.73, which is 0.17 higher than that of baseline, showing quite dramatic increment.

Table 1: Final results of experiments//Energy consumption during the training

	PSDS1	**PSDS2**
Flcam-CRNN	0.437 +/- 0.001	0.651 +/- 0.003
Mha-CRNN	0.069 +/- 0.001	0.730 +/- 0.003

Table 2: Energy consumption during the training and evaluation

	Training (kWh)
Flcam-CRNN	1.390 +/- 0.019
Mha-CRNN	1.418 +/- 0.016

5. REFERENCES

[1] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[2] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, 2019.

[3] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017.

[4] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*. ISCA, sep 2019.

[5] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[7] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.