SOUND EVENT DETECTION SYSTEM USING A MODIFIDED MVIT FOR DCASE 2023 CHALLENGE TASK 4B

Technical Report

Shutao Liu[1]*, Peihong Zhang*[2]*, Fulin Yang*[2]*,*
Chenyang Zhu[1]*, Shengchen Li*[2]*, Xi Shao*[1]

[1] Nanjing University of Posts and Telecommunications,
Nanjing, Jiangsu, P.R.China,
{1022010303,Zhuchenyang,Shaoxi}@njupt.edu.cn
[2] Xi'an Jiaotong-Liverpool University,
Suzhou,Jiangsu, P.R.China,
Shengchen.Li@xjtlu.edu.cn
{Peihong.Li20,Fulin.Yang20}@student.xjtlu.edu.cn

ABSTRACT

In this report, we describe our submissions for the task 4b of Detection and Classification of Acoustic Scenes and Events (DCASE) 2023 Challenge: Sound Event Detection with Soft Labels. We use a MViT model based on frequency dynamic convolution. While preserving the advantages of multi-scale feature extraction by MViT, frequency dynamic convolution is used to overcome the translation invariance of image feature extraction caused by MViT model to improve the ability of the model in terms of extracting frequency dimension features. Without using any external datasets or pretrain model, our system trained only on the provided soft-label dataset, and the final F1-m score and F1-MO score are 80.52 and 63.43, respectively, both higher than the baseline system.

Index Terms— DCASE, sound event detection, frequency dynamic convolution, multiscale vision transformer(MViT), soft labels

1. INTRODUCTION

The goal of sound event detection(SED) is to give the start and end time of sound events while classifying them. The main task of DCASE2023 task4b is to explore how to use soft labels with more distribution information to improve the effect of model on SED tasks [1].

In order to take full advantage of the data distribution information provided by soft labels, we consider to improve the accuracy of the model by utilizing the multi-scale feature extraction strategy. We adopted MViT model, and based on its multi-scale hierarchical model, we improve the ability of multi-scale features extraction [2]. Since sound event detection is dependent on the frequency dimension, frequency dynamic convolution is used to replace the convolutional layer in the model [3]. Our system achieves good results in sound event detection with soft labels.

2. PROSED METHORD

Soft label is a new type of label that has recently been proposed for use with SED tasks. It provides as a number between 0 and 1 that characterize the certainty of human annotators for the sound at that specific time[4]. Compared with hard labels, soft labels are considered to be a form of labels that can provide more data distribution information. The challenge of this competition is how to use these more detailed data distribution information to improve the accuracy of models in SED tasks. In order to take full advantage of soft label information for training, we propose a model whose idea is derived from the MViT model. Its can find the required information from a larger number of scales, but the MViT model is still slightly inadequate in coping with tasks of ascension such as sound time detection. Its application for shift invariance has a negative impact on the acoustic direction. So we adopted the method of frequency dynamic convolution to overcome its negative effects. Frequency dynamic convolution has been shown to be an extremely useful convolution method for acoustic directions and is widely used in sound time detection tasks.

2.1. MViT

The idea of MViT is similar to that of SWIN Transformer[5], in that both change the perceptual field by varying the patch size, hoping to extract different scales of information at different perceptual field sizes. pooling operation and Swin is a Windows operation.

MViT modifies the Transformer by adding the pooling operation after Q,K,V,just like the figure1 shows.

However, Transformer can only handle one-dimensional data, so the shape of the spectrogram signal after the patch processing becomes $(L, D), L = THW$ that is THW in the figure , Self-attention calculation formula is mainly QK^TV , assuming that:

$$Shape_Q = (L_Q, D)$$
$$Shape_K = (L_K, D)$$
$$Shape_V = (L_V, D)$$

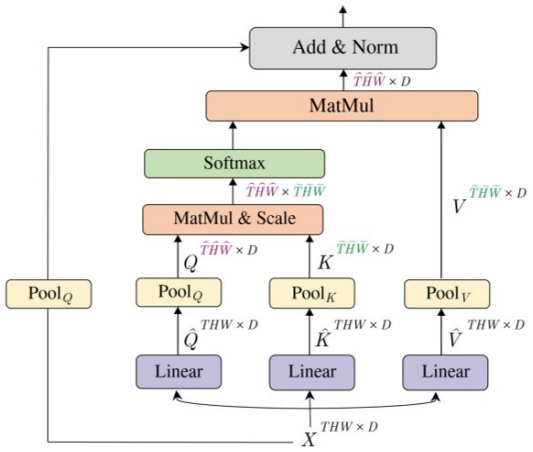

Figure 1: Specific structure of MViT block.

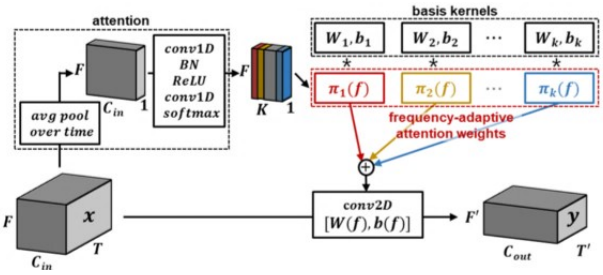

Figure 2: FDY.

then we will get:

$$Shape(QK^TV) = (L_Q, L_K)(L_V, L_D) = (L_Q, D)$$

in order to make the formula hold, we must ensure that $L_K = L_V$, that is THW, the green in Figure 1.

so in order to reduce the spatial resolution, only need to change the sequence length of the Q vector, so the Q vector pooling operation can be carried out, while experimental evidence that K,V vector pooling will improve the index, so the K,V vector is also pooling operation, but will not affect the size of the spatial resolution, in order to ensure that the res connection holds, need to The same pooling operation as Q vector is performed for input X.

The pooling operation is divided into max, average and conv, etc.

Previous experiments have demonstrated that using convolution to reduce the resolution is the best approach among these methods.

Therefore we may wish to make some improvements to the convolution itself to improve the performance of the MViT model itself.

2.2. Frequency Dynamic Convolution

For this, we used frequency dynamic convolution, which is widely used in SED tasks and has proven itself to be a good performer.

The frequency dynamic convolution was proposed to maintain translation equivariance of 2D convolution on time dimension while loosening it on frequency dimension to improve model's physical consistency with sound events' time-frequency patterns and to improve SED performance.

Frequency dynamic convolution uses frequency-adaptive kernel in order to enforce frequency-dependency on 2D convolution thus to improve physical consistency of SED model with sound events' time-frequency patterns. The operation is illustrated in Figure 2. It first extracts frequency-adaptive attention weights from input by applying average pooling over time axis followed by two 1D convolution layers along channel axis. Instead of using fully-connected (FC) layers as dynamic convolution did, we applied 1D convolution in order to consider adjacent frequency components as well. Between two 1D convolution layers, batch normalization and ReLU are applied. 1D convolution layers compress channel dimension into the number of basis kernels. Then, softmax is applied tomake frequency-adaptive attention weights range between zero and one and make sum of the weights for different basis kernel equal to one. Temperature of 31 was applied on the softmax to ensure uniform learning of basis kernels and stable training. Then frequency-adaptive convolution kernel is obtained by weighted sum of basis kernels using frequency-adaptive attention weights, where basis kernels are trainable parameters as well. Obtained frequency-adaptive kernel is used for frequency dynamic convolution operation just as normal 2D convolution.

3. EXPERIMENT

We use the MViT setup in our experiments. Due to computer performance problems, experiments are conducted on a small MViT model with 10 MViT blocks (denoted as Mvit-S) and a large MViT model with 24 MViT blocks (denoted as Mvit-B) respectively. It turns out that the MViT model has quite good performance, far exceeding the performance of the baseline model. The use of frequency dynamic convolution in turn builds on this.

We used 128 frequency bins for the Mel spectral transform of the audio, and the hope-size was set to 20ms. The input audio length is set to 175 or 3.5s, to fit the output length of 7.In other words, the time resolution is set to be 0.5s, and is sufficient for the final 1s resolution and can better resolve some unexpected disturbances, making the output more stable.

We also made some changes to the MViT code itself to make it able to accept inputs that are not square. We have further improved the performance of the model by inputting extended audio that contains a period of time before and after the audio, again using frequency dynamic convolution alone.

All the results are shown in the table below. We can find that the FDY works quietly well and the pad method can improve performance even a little bit more.

Model	F1-m	F1-MO
Baseline	71.50	35.21
MViT-B	74.45	59.56
MViT-B + FDY	**78.53**	**61.15**
MViT-S	69.13	53.24
MViT-S + FDY	72.91	56.43
MViT-B + FDY + Pad	**80.52**	**63.43**

4. CONTLUTION

In this report, we present our methods used in the task 4b of DCASE 2023 Challenge. We use MViT model combined with frequency dynamic convolution to improve the ability of the model in terms of extracting frequency dimension features when extracting multi-scale features. In addition, we also use padding method to further improve the performance of the model. Our final systems achieve a F1-m/F1-MO score of 0.8052/0.6343 on development dataset.

5. REFERENCES

[1] http://dcase.community/challenge2023/task-sound-event-detection-with-soft-labels.

[2] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," 2021.

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

[4] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.