MULTI-SCALE CONVOLUTION BASED ATTENTION NETWORK FOR SEMI-SUPERVISED SOUND EVENT DETECTION

Technical Report

Xiujuan Zhu[1,3], *Xinghao Sun*[1,3*], *Ying Hu*[1,3], *Yadong Chen*[1,3], *Wenbo Qiu*[1,3], *Yuwu Tang*[1,3], *Liang He*[1,2], *Minqiang Xu*[4]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, China
{xiujuanzhu841@gmail.com}
[2] Department of Electronic Engineering, Tsinghua University, China
[3] Key Laboratory of Signal Detection and Processing in Xinjiang, China
[4] SpeakIn Technology

ABSTRACT

Deep Convolutional Recurrent Neural Networks (CRNN) have drawn great attention in sound event detection (SED). Due to the variation in duration for acoustic events is relatively large, It is critically important to design a good operator that can extract multi-scale feature more efficiently for SED. However, most CRNN-based models lack discriminative ability for different types of acoustic events and deal with them equally, which results in the representational capacity of the models being limited. Inspired by this, We proposed a Multi-Scale Convolution based Attention Network(MSCA). By using Multi-Scale Convolution, a more effective feature representation ability can be obtained, Which can naturally learn coarse-to-fine multi-scale features to helps the model recognize different sound events. On the other hand, a channel-wise attention module is designed, which can adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels.

Index Terms— Multi-Scale, Channel-wise Attention, Sound event detection

1. INTRODUCTION

The purpose of Sound Event Detection is to identify each sound event category and detect its onset and offset in an audio sequence. Unlike common classification tasks, that is only need to determine the event category, while the detection task is need to predict the temporal position of the target events, which increased the difficulty of the SED task. SED has drawn great attention recently in a variety of applications, such as query-based sound retrieval[1], smart cities, and homes[2][3] , as well as multimedia information retrieval[4]. There main there approaches exist to train an SED model: Fully supervised SED and Weakly supervised SED and Semi-supervised SED. However, Following the dcase2021Task4, this paper primarily focuses on Semi-supervised SED based on mean teacher method.

Recently, due to the development of deep convolutional neural networks, there has been significant progress on the problem of sound event detection. Existing approaches for SED can be roughly classified into two kinds, i.e.,cnn framework[5][6] and crnn framework[7][8][9]. However, some works shown that CNN models generally good at audio tagging, while CRNN approaches are excel at in onset and offset detection. Thus, we design our model based on CRNN framework. On the other hand, Different sound events behave differently in the time and frequency domains. For example, dog barking and dish last shorter, while running water and blender lasts longer and cover a wider range in the frequency domain. If the model perform on a single resolution, It's hard to deal with different types of sound events. Thus, Consider this inherent nature of sound events, Inspired by [10], We design a multi-scale feature extraction module to help respond to different types of sound events. In order to fusion features of inconsistent semantics and scales, inspired by SENet[11], a channel-wise attention module is designed to better recognize sound events.

2. PROPOSED METHOD

In this section, an overview of the proposed framework is firstly illustrated. Then we introduce the proposed Multi-scale Convolution and Channel-wise attention model. Multi-Scale Convolution exploits different types of filters with varying size to integrate the multi-scale spatial information for better representation, while Channel-wise attention model aims to establish a cross-channel interaction in different scale features and select more important scale for better perception of different types of acoustic events.

2.1. proposed network

We adopt the CRNN Network as the basic network structure to explore the effects of the Multi-scale Convolution and the Channel-wise Attention model for the sound event detection. As show in Fig.1 the processing flow of our proposed MSISA Network mainly consists of three stages: the CNN feature extraction stage, the RNN long-time context modeling stage, and the localization output stage based linearsoftmax[12]. As illustrated in Fig.2, The CNN part mainly consist of several Multi-scale Convolution Attention Blocks(MSCA Block) following the average pooling operation . And a MSCA Block consists of two MSCA modules based on the residual connection.

*equal contribution with Xiujuan Zhu

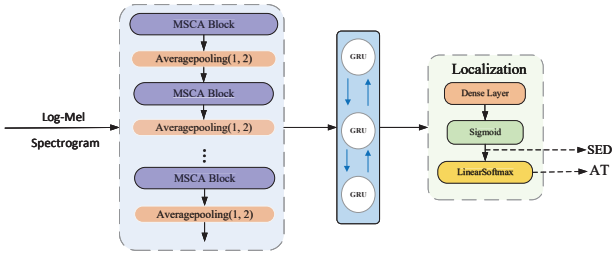

Figure 1: Overview of our architecture.

2.2. Multi-scale Convolution

The Multi-scale convolution can process the input feature at multiple scales in parallel, As illustrated in Fig.3, where a four-branch case is shown. Therefore in this example, there are only four kernels with different kernel sizes, but it is easy to extend to multiple branches case. For a given feature map $X \in R^{C\times H\times W}$, we first conduct four transformations with different kernel sizes k_i, and the input feature map X can be splitted into four parts as denoted by [X_0, X_1, X_2, X_3], For different parts, different spatial resolutions and depths can be generated by using multi-scale kernels in the structure. Thus the multi-scale feature map generation function can be writing as:

$$X_i = Conv(k_i \times k_i)(X), i = 0, 1, 2, 3 \tag{1}$$

where the k_i is the the i-th kernel size, $X_i \in R^{C\times H\times W}$ denotes the feature map with different scales and then the multi-scale preprocessed feature map can be obtained by a concatenation way as

$$F = Concat([X_0, X_1, X_2, X_3]) \tag{2}$$

where the $Concat$ means to concatenate features in the channel dimension. $F \in R^{4C\times H\times W}$ is the obtained multi-scale feature map.

2.3. Channel-wise attention module

By extracting the channel attention weight information from the multi-scale pre-processed feature map, the attention weight vectors with different scales are obtained. The vector of attention weight Z can be represented as

$$Z = Softmax(SEWeight(F)) \tag{3}$$

Where $Z \in R^{4C\times 1\times 1}$,

$$Y = Z \odot F \tag{4}$$

where $\odot$ represents the channel-wise multiplication, $Y \in R^{4C\times H\times W}$ refers to the feature map that with the obtained channel-wise attention weight. Then a 1x1 convolution is used to learn the correlation between channels. The final output of the model can be writting as $Y^{'} \in R^{C\times H\times W}$,

$$Y^{'} = Conv1 \times 1(Y) \tag{5}$$

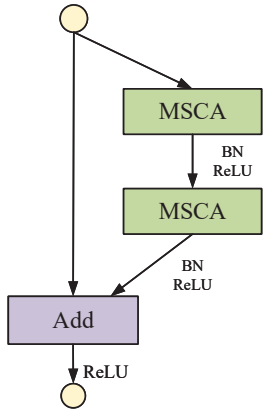

Figure 2: Multi-Scale Convolution based Attention block.

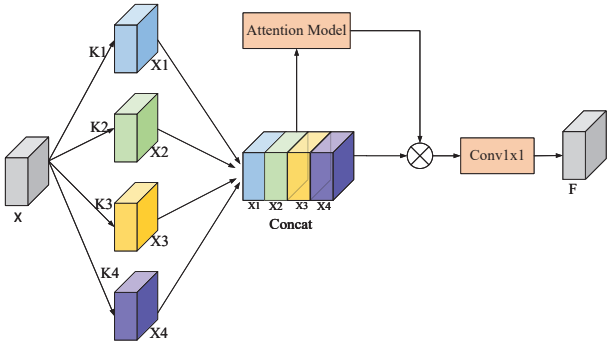

Figure 3: Multi-Scale Convolution based Attention model. K_i is the kernel size, X_i is the fearure map with different scales.

3. IMPLEMENTATION

The proposed model are a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN). In our experiment, We use three branches to implement the Multi-scale Convolution. The CNN part is composed by 7 layers. In the first four layers, the kernel size is [[3, 3], [5, 5], [7, 7]] , while in the latter three layers, the kernel size is set to [[3, 3], [5, 3], [7, 3]]. The number of filters and pooling size for each layer are respectively [16, 32, 64, 128, 128, 128, 128] and [[1, 2], [1, 2], [2, 2], [2, 2], [1, 2], [1, 2],[1, 2]]. In order to reduce the complexity of the model, we use different sizes of grouping convolution at different layers. The dropout is set to 0.3. The batch size is set to 48. The RNN part is composed by two layers of RNN cells, each layer contains 128 cells. And a aggregation layer (in our case a linearsoftmax) is used to aggregates frame-level features to segment-level features to produce audio tagging output.

4. EXPERIMENTS

4.1. Dataset

The dataset of dcase2021 task 4 is consist of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. Three types dataset (i.e. the weakly labeled data(1578), unlabeled data(11412) and synthetic data(10000)) were

used for training in each batch as proposed in the baseline (1/4 for the weakly labeled data, 1/2 for the unlabeled data and 1/4 for the synthetic data). In order to learn efficiently with unbalanced training set, we use the mean-teacher model that is provided by the baseline system.

4.2. Results and Analysis

In DCASE 2021 task 4, the PSDS-scenario1 and the PSDS-scenario2 is used to evaluate the performance. The results that we obtained for our proposed models are given for the validation dataset(1168).

5. CONCLUSION

In this technical report, we present a system for DCASE2021 task 4. We proposed a Multi-Scale Convolution based Attention Network for sound event detection. The Multi-Scale Convolution module can fully extract the different-scale event features by using the different kernels, while channel-wise attention model can adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. An PSDS-scenario1 of 0.342 and PSDS-scenario2 of 0.614 was achieved on the validation data. Due to lack of time, there are still potential improvements can be achieved by adjusting the parameters in this model. Thus, we will carry out a further work in the model.

References

[1] F. Font, G. Roma, and X. Serra, "Sound sharing and retrieval," in *Computational analysis of sound scenes and events*. Springer, 2018, pp. 279–301.

[2] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 373–397.

[3] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[4] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[6] Y. Huang, X. Wang, L. Lin, H. Liu, and Y. Qian, "Multi-branch learning for weakly-labeled sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 641–645.

[7] W. Ding and L. He, "Adaptive multi-scale detection of acoustic events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 294–306, 2019.

[8] X. Zheng, Y. Song, J. Yan, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," *Proc. Interspeech 2020*, pp. 841–845, 2020.

[9] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021.

[10] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: rethinking convolutional neural networks for visual recognition," *arXiv preprint arXiv:2006.11538*, 2020.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.