Multi-scale Convolutional Recurrent Neural Network and Data Augmentation for Polyphonic Sound Event Detection

Keming Zhang¹, Yuanwen Cai¹, Yuan Ren¹*, Ruida Ye¹, Xianwei Zhang², Liang He²*

¹Space Engineering University, Beijing, 101416, China
²Department of Electronic Engineering, and Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China

*Corresponding author’s email: renyuan_823@aliyun.com, heliang@mail.tsinghua.edu.cn

Abstract. We propose Multi-scale convolutional recurrent neural networks (MCRNN) and data augmentation methods to detect polyphonic sound event with few training data. MCRNN consists of Multi-scale convolutional neural networks (MCNN) and recurrent neural networks (RNN). MCNN concatenates the higher level features extracted using multiple convolution kernels with different scales from the time domain and frequency domain at the same time. RNN is able to capture the longer term temporal context characteristics. A novel background spectrum random replacement (BSRR) data augmentation method is applied to expand training data, which uses standard normal distribution data with randomly selected position and length instead of the original time-domain, frequency-domain or time-frequency domain background spectrum features. Our method is tested on the datasets of DCASE 2019 Task3 (T3). The experimental results showed that the MCRNN and BSRR data augmentation method are efficient. We achieved better results than the first place and the single advanced on the T3 by applying BSRR and SpecAugment data augmentation method simultaneously. On the evaluation dataset (T3-eval), our best result shows 0.05 and 0.975 of error rate (ER) and F1 respectively. Our method got the best performance and relatively improved 17% and 1% than the corresponding values of the single advanced.

Keywords. pattern recognition; sound event detection; data augmentation; Multi-scale learning; convolutional recurrent neural network

1. Introduction

Sound event is the sound record that humans labeled as a distinctive event in the continuous acoustic signal [1]. Sound Event Detection (SED) is the process to recognize various types of specific sounds from an audio clip and locates the presence timestamp [2]. Sound Event Detection has recently attracted significant attention in artificial intelligence research, which is used for multimedia, transportation, security monitoring and health care etc.

There are many different approaches for Sound Event Detection. Gaussian Mixture Models (GMM) [3], Hidden Markov Model (HMM) [4], Nonnegative Matrix Factorization (NMF) [5], or Support Vector Machine (SVM)[6] used for SED in history. However, their capability is still quite
limited. Deep learning has been applied to SED recently. Deep Neural Networks (DNN) [7], Convolutional Neural Networks (CNN) [8], Recurrent Neural Network (RNN) [9] and various combinations get rapid development. [10] described a sound event classification method using DNN which is better than Support Vector Machine (SVM). [11] proposed Convolutional Neural Networks (CNN) architectures to classify the sound event. [12] introduced a multi-scale time-frequency attention module for rare audio event detection. These method mainly focused on monophonic audio events detection or classification.

In the complex environment, multiple audio events may alternately occur or overlap at the same time. Polyphonic sound event detection are much more important and challenging than the monophonic detection. [13] presented a method based on bi-directional long short term memory (BLSTM) for polyphonic sound event detection. Çakir combined CNN and RNN to detect polyphonic sound event, and achieved a considerable performance improvement than CNN, RNN, and other established methods. [14] proposed a consecutive ensemble of CRNN models, which won the first place on Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 sound event localization and detection task.

The rest of the paper is organized as follows. Section 2 describes the proposed method. The test experiments in detail and results analysis on the TUT Rare Sound Events 2019 dataset are presented in Section 3. Section 4 concludes and discusses future work.

2. Method

This section describes the method proposed for polyphonic SED. We present a novel data augmentation approach to generate more training data. We further introduce a new Multi-scale convolutional recurrent neural networks to detect the polyphonic sound event which includes Multi-scale convolutional neural networks (MCNN), recurrent neural networks (RNN) and fully connected (FC) layer.

2.1. BSRR data augmentation

Data augmentation has been proposed to generate additional training samples, which have achieved excellent performance in the image domain. Inspired by the SpecAugment [15], we construct a new background spectrum random replacement (BSRR) data augmentation method which is used to generate a large number of training samples with rich and varied spectrum characteristics to solve the problem of insufficient training data. We view the input log mel spectrum as spectrogram image features where the time axis is horizontal and the frequency axis is vertical. Given a multi-channel signal, we suppose the input features spectrogram has a size of \((T \times F)\), where \(T\) and \(F\) correspond to the input length along time and frequency axis respectively. We use randomly generated standard normal distribution data to replace the background sound events log mel spectrum feature data. We adapt three policies to increase the training samples, which include time-domain background spectrum random replacement (TBSRR), frequency-domain background spectrum random replacement (FBSRR), and time-frequency-domain background spectrum random replacement (TFBSRR). Figure 1 shows examples of the individual augmentation applied to the four channels training data.

TBSRR: TBSRR data augmentation policy is applied that \(\tau \times D\) consecutive background spectrum \([t_{bi}, t_{bi} + \tau, D]\) are replaced by standard normal distribution data for each log mel spectrum channels, where \(\tau\) is chosen from a uniform distribution from 0 to \(T\), \(t_{bi}\) is the start position where randomly selected along the time-domain direction of the background audio log mel spectrum. For examples, given a 3000×128 log mel spectrum feature, we randomly select a position along the time-domain direction in each channel, then use the random function to generate \(\tau \times 128\) size standard normal distribution data instead of the original background spectrum feature data. According to the needs, several times more time-domain augmentation data than the training data are generated.
Figure 1. Examples of BSRR data augmentation method applied to the base input log mel spectrogram. From top to bottom, the figures depict the log mel spectrogram of the base input training data with no augmentation, TBSRR, FBSRR and TFBSRR applied. The white rectangular denotes the position where TBSRR data augmentation method is applied, while the yellow rectangular labeled the position where FBSRR data augmentation method is applied.

FBSRR: FBSRR data augmentation policy is applied that \( t_b \times f \) consecutive background spectrum \([t_b, f_{bi}:f_{bi}+f]\) are replaced by standard normal distribution data for each log mel spectrum channels, where \( t_b \) means the frame length of the background audio log mel spectrum, \( f \) changed from 0 to \( F \) randomly, and \( f_{bi} \) is the start position randomly selected between 0 and \( f \) along the frequency domain direction. Given a \( 3000 \times 128 \) log mel spectrum feature, we randomly select a position along the frequency domain direction in each channel, then use the random function to generate \( t_b \times f \) size standard normal distribution data instead of the background spectrum feature data. According to the needs, several times more frequency domain augmentation data than the training data are generated.

TFBSRR: Using the training log mel spectrum feature data as the base input, TBSRR and FBSRR data augmentation policies are applied at the same time. Several times more augmentation data than the training data are generated.

2.2. MCRNN
Multi-Scale acoustic event detection was proposed by Zhang et al. in [16] which uses the Multi-scale hourglass structure. However, it's performance relies on the amount of annotated data and the convolutional kernels are fixed. Our proposed MCRNN Multi-scale method uses three groups of parallel CNN with different convolutional kernels to get more features from the input, and improves the time-frequency feature extraction capability for few training data. The MCRNN network
architecture is composed of MCNN, RNN and fully connected layer. The MCRNN architecture is shown in Figure 2.

![Figure 2. Example of BSRR data augmentation method on log Mel spectrogram.](image)

2.2.1. Multi-scale Convolutional Neural Networks
MCNN is made up of three groups of parallel CNN to get features from the input. Each group is composed of three layers convolutional modules with different convolutional kernels. The convolutional module consists of CNN, Batch normalization (BN), rectified linear units (ReLU), and Dropout. CNN includes 3×3, 5×5 and 7×7 convolutional kernels to get both the frequency-domain features and the time-domain features. Batch normalization (BN) and rectified linear units (ReLU) activation function is followed after each convolutional kernel. The dropout rate of 0.2 is employed for avoiding overfitting. Because of the shortest duration of the target audio event samples is about 0.2 second, we adopted 60 output features for the last 3×3 and 5×5 convolutional kernels to obtain fine features, and 8 output features for the last 7×7 convolutional kernels to extract global information. These features are concatenated to 128 output features.

2.2.2. RNN
RNN is good at processing sequence information related to context and has now become one of the most important audio recognition tool. The Gated Recurrent Unit(GRU) is a common RNN which can capture the context characteristics of time series effectively. GRU is usually used for training with inadequate data, which has fewer parameters. We use a Gated Recurrent Unit which has 64 hidden units and two layers. In the training process, we adopt 20% dropout rate to prevent over-fitting.

2.2.3. Fully connected layer
We apply a fully connected layer to output polyphonic sound events probability for each time frame. Activation function ReLU is used to activate the output. We adopt threshold value 0.5 to decide the polyphonic sound events occur or not. At last, the polyphonic sound events occurrence time (onset) and the disappearance time (offset) can be obtained.

3. Experiments and analysis
In order to assess the performance of our proposed method, we conducted experiments on the IEEE DCASE 2019 Challenge Task 3 dataset(T3). The network was trained using the adam algorithm for gradient-based optimization[17]. Training was performed for a maximum of 200 epochs using a learning rate of 0.001. We utilize early stopping criteria, and 10 epochs patience to train the model.

3.1. Dataset
We evaluate the proposed method on DCASE2019 T3 which provides 4 channels First-Order Ambisonic recordings. The dataset include development dataset(T3-dev) and evaluation dataset(T3-eval)[18]. The T3-dev consists of 400 recordings with 60-second long. The T3-eval consists of 100 recordings. Each recording contains either polyphonic sound events, or monophonic sound events[19].
On the T3-dev, we use 4-fold cross-validation to learn more effective information from the limited dataset. We take turns to use two of the four cross-validation split subsets for training, one split for validation and the last one for the test. Each split has only a 1.67-hour recording, so the training data are very few. We calculate each split performance and report the average of them. On the T3-eval, we use 75% of the T3-dev to train our network, and the remaining 25% dataset for validation.

3.2. Results and analysis

3.2.1. Performance comparison to different augmentation data amounts
In order to compare polyphonic SED performance of different augmentation data amounts, we evaluated the performance with the original training data and the data after three to nine times augmentation data which is generated with our proposed BSRR data augmentation method. The average F1 and ER on the T3-dev, the best F1 and ER on the T3-eval are present in Table 1. The results show that the performance is getting better with three to six times augmentation data on both T3-dev and T3-eval. The performance degrades with 9 times augmentation data. Using the BSRR data augmentation method on the T3-dev, we achieve the best ER 0.12, a relative 33% reduction over the original data, and the best F1 0.922, a 4% relative improvement to the original training data. The best ER/F1 on the T3-eval is 0.06/0.967, which performance improves 40% and 2% over the original data relatively. The results show that the data augmentation method improves the SED performance, but too many augmentation data will not improve performance further, the appropriate is the best.

Table 1. ER and F1 of different augmentation data on the DCASE 2019 T3

| Data   | T3-dev | T3-eval |
|--------|--------|---------|
|        | ER     | F1      | ER     | F1     |
| Original | 0.18   | 0.884   | 0.10   | 0.947  |
| 3 times | 0.13   | 0.918   | 0.09   | 0.950  |
| 6 times | 0.12   | 0.922   | 0.06   | 0.967  |
| 9 times | 0.14   | 0.914   | 0.07   | 0.958  |

In order to generate a large number of training samples with rich and varied spectrum characteristics, we expand the SpecAugment method which applied zero to log mel spectrogram frequency and time masks, and implement the GFN and RFI sound data augmentation methods which use random numbers instead of the spectrum feature data. GRN sound data augmentation method replaces the log mel spectrum training data on the DCASE 2019 T3 with Gaussian random noise randomly. RFI sound data augmentation method uses random floats instead of the original training data log mel spectrum. The BSRR data augmentation performance is compared with SpecAugment, GRN and RFI and the combination of them. The F1/ER of different data augmentation methods and the combination of them on the DCASE 2019 T3 are shown in Table 2. Six times augmentation data is applied for each method. From the results, we conclude that the performance of SpecAugment data augmentation is better than other single data augmentation method. We used both BSRR and RFI data augmentation methods at the same time, the performance is slightly improved, the F1 on the T3-eval is worse than BSRR. We utilized the combination of BSRR, SpecAugment, RFI three policies, which can get better performance than the single method. We achieve the best polyphonic SED performance by the combination of BSRR and SpecAugment data augmentation methods. We prove that applying different mixture data augmentation methods can improve the polyphonic SED performance.

Table 2. ER and F1 of different data augmentation method on the DCASE 2019 T3
Method | T3-dev | T3-eval
---|---|---
| ER | F1 | ER | F1 |
RI | 0.13 | 0.916 | 0.07 | 0.958 |
GRN | 0.14 | 0.917 | 0.06 | 0.964 |
SpecAugment | 0.11 | 0.932 | 0.06 | 0.966 |
BSRR | 0.12 | 0.922 | 0.06 | 0.967 |
BSRR+RFI | 0.11 | 0.931 | 0.06 | 0.963 |
BSRR+SpecAugment | 0.1 | 0.941 | 0.05 | 0.975 |
BSRR+SpecAugment+RFI | 0.1 | 0.939 | 0.05 | 0.970 |

3.2.2. Comparison to different network architectures on the DCASE 2019 T3
We tested different polyphonic sound event detection networks on the DCASE 2019 T3. The different network architecture parameters are shown in Table 3. In the Table, CRNN1 have three layers 3×3 convolution kernel. CRNN2 uses 3×3, 5×5 and 7×7 convolution kernels. MCRNN1 consists of two groups of parallel convolution kernel(3×3 and 5×5). MCRNN2 is composed of four groups of parallel convolution kernel(3×3, 5×5, 7×7, and 9×9). The same RNN and fully connected (FC) output are applied after convolutional layers of each different network.

| Network | Parallel | Network description |
|---|---|---|
| CRNN1 | no | 3×3conv-3×3conv-3×3conv-RNN-FC |
| CRNN2 | no | 3×3conv-5×5conv-7×7conv-RNN-FC |
| MCRNN1 | 2 | 3×3conv-3×3conv-3×3conv-3×3conv-5×5conv-5×5conv-5×5conv-RNN-FC |
| MCRNN2 | 4 | 3×3conv-3×3conv-3×3conv-3×3conv-5×5conv-5×5conv-5×5conv-7×7conv-7×7conv-7×7conv-9×9conv-9×9conv-9×9conv-RNN-FC |
| Our method | 3 | 3×3conv-3×3conv-3×3conv-3×3conv-5×5conv-5×5conv-5×5conv-7×7conv-7×7conv-7×7conv-RNN-FC |

3.2.3. Comparison with other methods
Table 5 shows the ER and F1 results of MCRNN, and the comparison with other polyphonic AED methods on DCASE 2019 T3. The baseline [20] method employ CRNN to generate benchmark scores on the DCASE2019 T3 datasets. CE-CRNN [21] won the first place on the T3 using consecutive ensemble of CRNN models. The single advanced method uses data augmentation, network prediction
and post-processing to get the state-of-the-art detection performance on the DCASE 2019 challenge T3 datasets[22]. From the results, we can see that our method achieved the best detection performance both on the T3-dev and T3-eval. On the T3-dev, we achieve the lowest ER 0.1, a 29% reduction over the single advanced, and the best F1 0.941, a 3% performance improvement relative to the single advanced. The best ER/F1 on the T3-eval is 0.05/0.975, which performance improves 17%/1% over the single advanced. Compared to the 1st place, the F1 of our method improved by 3%, and the ER reduces by 38% relatively.

| Method              | T3-dev | T3-eval |
|---------------------|--------|---------|
|                     | ER     | F1      | ER     | F1      |
| Baseline[48]        | 0.3    | 0.799   | 0.28   | 0.85    |
| CE-CRNN[49]         | 0.1    | 0.893   | 0.08   | 0.94    |
| Single advanced[50] | 0.1    | 0.916   | 0.06   | 0.96    |
| Our method          | 0.1    | 0.941   | 0.05   | 0.97    |

4. Conclusions
We proposed Multi-scale convolutional recurrent neural networks (MCRNN) and novel background spectrum random replacement (BSRR) data augmentation methods which are efficient to detect polyphonic sound events with few training data. The proposed BSRR data augmentation method is used to expand training data. Experiments show that the detection performance can be improved by using the appropriate augmentation data amounts and various data augmentation methods alone or in combination. We present a MCRNN method which uses three groups of parallel concatenated Multi-scale CNN to extract high-level features from the input log mel spectrum and a RNN to capture context information, which is better than other CRNN and MCRNN methods that we tested on the DCASE 2019 T3 with few training data. We achieved the best polyphonic SED performance on T3 by using both BSRR and SpecAugment data augmentation methods to increase the augmentation data by 6 times respectively. In the future, we will continue to test our method with more data augmentation methods and other datasets, strive to improve the polyphonic SED performance and put it into engineering practice. We share all the python code on the GitHub, where it can be downloaded at https://github.com/zhang201882/MCRNN.

5. References
[1] E. Çakir, T. Heittola, H. Huttunenc and T. Virtanen 2015 Polyphonic sound event detection using multi label deep neural networks.2015 international joint conference on neural networks (IJCNN). Killarney Convention Centre in Killarney, Ireland, pp. 1-7.
[2] H. Toni, M. Annamaria, V. Tuomas, and G. Moncef 2013 Supervised model training for overlapping sound events based on unsupervised source separation, in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 8677–8681.
[3] J. H. Wright 2010 Gmm with weak identification, Econometrica, vol. 68, no. 5, pp. 1055–1096.
[4] L. R. BAHL 1986 Maximum mutual information estimation of hidden markov model parameters for speech recognition, Proccasp.
[5] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste et al 2013 An exemplar-based nmf approach to audio event detection, in 2013 IEEE workshop on applications of signal processing to audio and acoustics. IEEE, pp. 1–4.
[6] L. Li, G. Fengpei, Z. Qingwei, and Y. Yonghong 2010 A svm-based audio event detection system, in International Conference on Electrical & Control Engineering.
[7]. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath 2012 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97.

[8]. I. M. H. Zhang and Y. Song 2015 Robust sound event recognition using convolutional neural networks, in icassp IEEE International Conference on Acoustics.

[9]. G. Alex, M. Abdel-rahman, and H. Geoffrey 2013 Speech recognition with deep recurrent neural networks.

[10]. I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao 2015 Robust sound event classification using deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 540–552.

[11]. S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al 2017 Cnn architectures for large-scale audio classification, in 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, pp. 131–135.

[12]. Z. JingYang, D. WenHao, and H. e. a. Liang 2019 Multi-scale time frequency attention for rare sound event detection, pp.3855–3859.

[13]. G. Parascandolo, H. Huttunen, and T. Virtanen 2016 Recurrent neural networks for polyphonic sound event detection in real life recordings, in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6440–6444.

[14]. K. Slawomir and L. Mateusz 2019 Sound source detection, localization and classification using consecutive ensemble of CRNN models, Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019.

[15]. S. P. Daniel, C. William, Z. Yu, C. Chung-Cheng, Z. Barret, D. C. Ekin, and V. L. Quoc 2019 Specaugment: A simple data augmentation method for automatic speech recognition, in Proc. 2019Interspeech, pp. 2613–2617.

[16]. J.Y. Zhang, W.H. Ding, and L. He, et al 2019 Multi-scale time-frequency attention for rare sound event detection, Interspeech 2019, Graz, Austria,2019, 3855–3859.

[17]. D. Kingma and J. Ba 2014 Adam: A method for stochastic optimization, Computer Science.

[18]. A. Shara th, P. Archontis, and V. Tuomas 2019 A multi-room reverberant dataset for sound event localization and detection, In Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), URL: https://arxiv.org/abs/1905.08346.2019.

[19]. G. Forman and M. Scholz, Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement, SIGKDD Explor. Newsl., vol. 12, no. 1, pp. 49–57, November 2010. [Online].Available: http://doi.acm.org/10.1145/1882471.1882479.

[20]. S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, IEEE Journal of Selected Topics in Signal Processing, pp. 1–1, 2018.

[21]. K. Slawomir and L. Mateusz 2019 Sound source detection, localization and classification using consecutive ensemble of CRNN models, Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).

[22]. L. H. Jingyang Zhang, Wenhao Ding 2019 Data augmentation and prior knowledge-based regularization for sound event localization and detection, Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).