SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS USING DENSE RECURRENT NEURAL NETWORK

Technical Report

Tianchu Yao[1], *Chuang Shi*[2], *Huiyong Li*[3]

University of Electronic Science and Technology of China, Chengdu, China
[1] tianchuyao@std.uestc.edu.cn
[2] shichuang@uestc.edu.cn
[3] hyli@uestc.edu.cn

ABSTRACT

In this paper, we introduce our sound events detection system using a mean-teacher model with convolutional recurrent neural network (CRNN) for DCASE 2020 Task4, which include residual convolutional block and dense recurrent neural network (DRNN) block. To improve the performance of system, we propose to use various methods such as multi-scale input layer, data augmentation, median window filters and model fusion. By combining those method, our system achieves 15% improvement on macro-averaged F-score on the development set, as compared to the baseline.

Index Terms— DCASE 2020, sound event detection, CRNN, mean-teacher model

1. INTRODUCTION

DCASE challenge is short for the challenge on detection and classification of acoustic scenes and events. Task 4 of DCASE 2020 challenge aims at detecting event with both weakly and strongly labeled data, together with a large amount of unlabeled data. In this task, the target of the sound event detection (SED) system is to provide the event class and the event time boundaries, given that multiple events can be present in an audio recording. In DCASE 2019, in order to improve the performance of SED systems, most of the researchers have proposed to use more complex networks than the baseline [1]. The most common way is to increase the number of channels and the number of convolution layers in feature encoders [2][3][4][5]. However, a rather simple decision layer is often adopted to analyze the output feature vectors from feature encoders and get the final results. The typical classifiers, such as the recurrent neural network (RNN) with linear layer, limit the analytical ability of SED systems in time axis [1][3][4][5].

In recent years, with the development of deep learning, researchers propose many powerful deep learning models. The dense convolutional network (DenseNet) has achieved significant performance in various tasks [6]. By creating the short-cutting paths linking early layers to later layers, features are shared and reused in different layers of the DenseNet. This can help the DenseNet to extract more robust feature representations.

In this report, we introduce our SED system designed for task 4 of DCASE 2020 challenge, including a mean-teacher model and CRNN, in which the mean-teacher model is used for utilizing the unlabeled training set. Firstly, we extract log mel spectrograms as the input feature and carry out data augmentation to prevent overfitting in the training process. An encoder, which consists of several residual blocks, is then used to extract feature from log mel spectrograms. Moreover, to make the system have stronger recognition ability in time dim, inspired by the DenseNet, we propose the DRNN as the decision layer of our system. Finally, we ensemble trained models to get the final SED results.

2. DATASET AND AUDIO PREPROCESSING

2.1. Dataset

Task 4 of DCASE 2020 challenge provides the training dataset named DESED [8]. This dataset consists of three parts: weakly labeled data, strong labeled data and unlabeled data. This task involves 10 class of sound events. Weakly labeled dataset contains 1578 clips and 2244 sound events occurrences. Unlabeled dataset contains 14412 clips. We use the pre-generated strong labeled dataset containing 2595 clips.

2.2. Audio Preprocessing

Our SED system works on data sampled by 16 kHz. Therefore, all the audio clips are resampled to 16 kHz at first. Then, we extract the log mel spectrogram feature from the audio files with 128 mel bands and the max frequency of 8 kHz. A 2048-point hanning window with the hop size of 256 is adopted. The features have the shape of 684 by 128. Thereafter, they were normalized by their mean and standard deviation.

2.3. Data Augmentation

Data augmentation is helpful to generate more training data, which can greatly improve generalization ability of the model and overcome the overfitting in the training process. In this report, we consider two data augmentation method: mixup [9] and time shift.

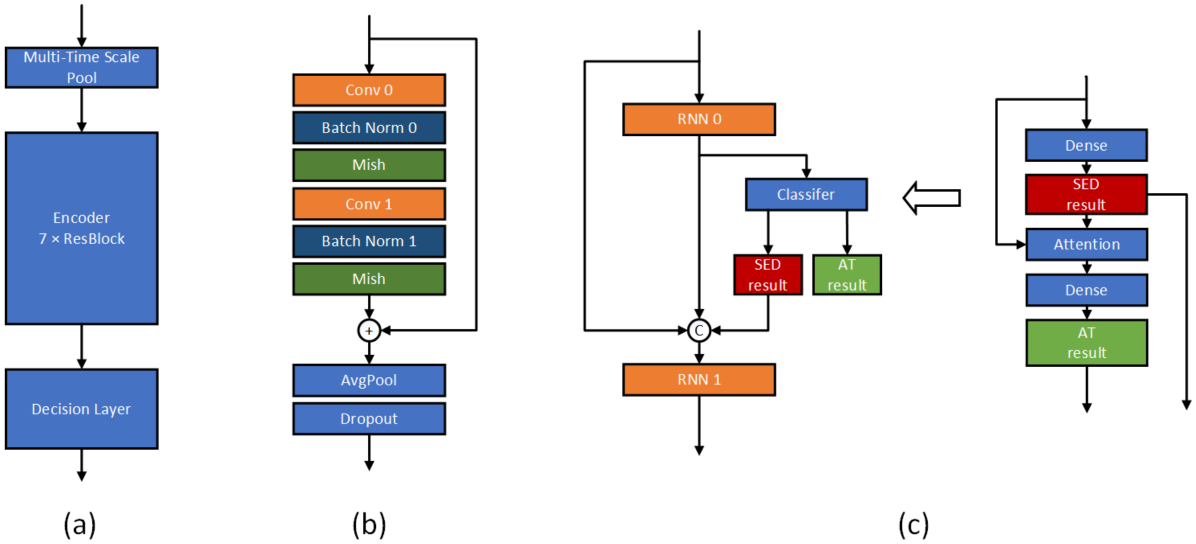

Figure 1 The structure of the proposed network. As shown in Fig.1.a, our network consists of a multi-time scale pool layer, an encoder based on several residual blocks and a decision layer. Fig.1.b is the illustration of residual blocks' structure. The structure of DRNN-2 is shown in Fig.1.c, in which the SED result and AT result are the predictions of SED and audio tagging respectively and the 'C' mean splicing operation. The input of RNN 1 consist of three parts: the input of RNN 0, output of RNN 0 and the SED result predicted by RNN 0.

2.3.1. Mixup

Mixup can improve the performance of deep neural network in many machine learning tasks by smoothing the distribution of samples in the feature space. This method creates a new data by interpolate between two raw data, while the labels are interpolated in the same way. This process is expressed as

$$\begin{aligned} \tilde{x} &= \gamma x_i + (1-\gamma) x_j \\ \tilde{y} &= \gamma y_i + (1-\gamma) y_j \end{aligned}, \qquad (1)$$

where x_i and x_j is two random chosen features, y_i and y_j is corresponding label respectively. γ is a random variable which follows the Beta distribution $\mathrm{Be}(\alpha, \alpha)$. In our system, we used $\alpha = 0.2$.

2.3.2. Time Shift

In the time shift method, raw mel spectrogram is shifted along the time axis for random frames. In our system, the time shifting frames follow uniform distribution from 0 to 684 frames. For samples whose strong labels are available, the strong labels will be shifted for same time shifting frames as corresponding spectrogram in synchronization.

3. MODEL DETAILS

CRNN has been proven to be efficient for SED tasks. We propose a mean-teacher model with CRNN, as shown in Fig.1. At first, we add a layer before CRNN called multi-scale input layer to get multi-time scale input features. Using residual block, the encoder layer of our network is designed deeper than the baseline to help the network to extract more robust feature representations of sound events. In decision layer, we propose the DRNN to detect sound events from the output feature vector of encoder layer. Finally, dense layers with attention pool is used for getting the results of SED and audio tagging which are the results in frame level and clip level respectively. The audio tagging results are used for the masking of SED result which enforce the SED results to focus only on the classes in audio tagging results.

3.1. Mean-Teacher Model

To utilize the unlabeled training set, we use the mean-teacher model. The mean-teacher model consists of a teacher network and a student network with same structure. In every training step, after the student network be updated, the teacher network is updated using the moving average of the student model's weights.
There are two kinds of loss to calculate out in a training step: consistency cost and classification cost. The cost for consistency between the predictions of teacher and student model is applied on unlabeled data, while the cost of classification is applied on weakly labeled data and strong labeled data.

3.2. Multi-Scale Input Layer

The durations of the sound events in different classes differ greatly. The durations of some sound events classes can be as long as 5-10 seconds which called background sound classes, while other classes with short durations called foreground sound classes. In a short time, background sounds are more stationary than foreground sounds. Hence, in the short time scale, model percepts foreground sound well, while in long time scale model percepts background sounds well. In consideration of this, we use a multi-time scale layer based on several average pools in time axis to get multi-time scale feature. In our work, 3 and 5 are used for average pool sizes.

3.3. Encoder Layer

Residual learning structure [10] is widely used to solve the degradation problem in deep neural networks. Our encoder layer consists of 7 residual blocks. Their channels are [16, 32, 64, 128, 128, 128, 128] and the pool sizes are [(2,2), (2,2), (1,2), (1,2), (1,2), (1,2), (1,1)]. The kernel is 3×3 on all layers, and the Mish [14] is used for activation function. We set the dropout rate is 0.5. The structure of residual block is shown in Fig.1.b. Compared with the residual blocks in [10], in our network, we change the position of activation function in residual block.

3.4. Decision Layer

To improve the robustness of decision layer in SED, inspired by the DenseNet, we propose DRNN. Like DenseNet, there are dense connections between every two RNN layers in DRNN. As an example, in DRNN-2 which include two RNN layers called RNN-0 and RNN-1, shown in Fig.1.c. Suppose the input of DRNN-2 is y_0, y_1 is the output of RNN-1 and s_1 is the SED result based on y_1. In the next, for RNN-1, its input is the splicing of y_0, y_1 and s_1. By this way, like DenseNet, there are dense connections between RNN-0 and RNN-1 in DRNN. On the one hand, using the dense connections, the later RNN layers make decisions based on the decisions of early RNN layers. On the other hand, RNN layers of DRNN can share the features with togethers which is in different abstract degrees by the dense connections. Using DRNN, the decision layer will think and make decision twice or more which will help to make more robust detections. In our work, we use DRNN with two different depths called DRNN-2 and DRNN-3, and the DRNN-2 including two RNN layer with 128 and 394 cells, while DRNN-3 consist of three RNN layers with 128, 394 and 394 cells.

We use a classifier based on dense layers with attention same as [1], shown in Fig.1.c, to get the SED results and audio tagging results.

3.5. Model Fusion

A reliable approach to improve the performance of deep learning model is to have an ensemble of several trained models. This method combines several trained models to create a stronger model with better generalization performance. We use geometric mean to make a final ensemble decision.

3.6. Median Window

The SED output is postprocessed by median filtering, by which the SED predictions will be more smoothed to reduce the frame-level errors in detection results. The size of each event class was calculated from the statistic of the synthetic data.

4. EXPERIMENT AND RESULTS

4.1. Training

For evaluating the performance of the proposed methods, the dataset and audio preprocessing described in chapter 2 is used. The Adam optimizer [11] is used for training, and the learning rate decay from 1e-3 to 1e-7 with cosine annealing [12] is used for learning rate schedule. The binary cross-entropy function is used as the loss function of classification loss, while the mean square error is used as the consistency loss.

We trained several models which have different complexity of their decision layer. The decision layer of Model Ⅰ is a single RNN, while the decision layers of Model Ⅱ and Model Ⅲ are DRNN-2 and DRNN-3 respectively. In DRNN training, we use transfer learning method to training the RNN layers of DRNN one by one, take DRNN-2 as an example, at first we train the RNN-0 in DRNN-2 and encoder layer for 512 epochs, then we freeze the RNN-0 and the encoder layer and train RNN-1 for 128 epochs. By this training method, the later RNN layers can make decision based on the more reliable decisions of early RNN layers.

4.2. Evaluation and Results

In evaluation, we ensemble several sub-models in different training epochs of Model Ⅰ, Model Ⅱ and Model Ⅲ and get Fusion-Model Ⅰ, Ⅱ and Ⅲ. Moreover, we ensemble Fusion-Model Ⅰ, Ⅱ and Ⅲ and get the Fusion-Model Ⅳ. The submissions will be ranked according to the event-based F1-score computed over the real recordings in the evaluation set, thus we focus on the event-based F1-score of our models. Table 1 shows the results on validation set, with the event-based F1-score is used to evaluate the performance. These results indicate that the proposed method outperforms the baseline scores.

Table 1 Class-wise event-based F1-score of our systems

Model	Event-based F1-score
Baseline [1]	34.8
Fusion-Model Ⅰ	47.7
Fusion-Model Ⅱ	49.5
Fusion-Model Ⅲ	49.7
Fusion-Model Ⅳ	48.8

5. CONCLUSION

This paper presented a network structure named DRNN in SED system for the DCASE 2020 task 4. The DRNN consist of several RNN layers with the dense connections between every RNN layers. By using dense connections, the RNN layers of DRNN can share feature with together. Using the DRNN to be decision layer, the SED system improves its recognition ability for complex sequence features and make more precise prediction benefit from it. Finally, our models reach 49.7% of event-based F1-score.

6. REFERENCES

[1] L. L. Delphin-Poulat and C. Plapous, “Mean teacher with data augmentation for DCASE 2019 task 4,” in *the 2019 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2019.

[2] L. Lin, X. Wang, H. Liu and Y. Qian, “Guided learning convolution system for DCASE 2019 task 4,” in *the 2019 IEEE*

AASP Challenge on Detection and Classification of Acoustic Scenes and Event, 2019.

[3] W. Lim, S. Suh, S. Park and Y. Jeong, “Sound event detection in domestic environments using ensemble of convolutional recurrent neural networks,” in *the 2019 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Event*, 2019.

[4] J. Yan and Y. Song, “Weakly labeled sound event detection with resdual crnn using semi-supervised method,” in *the 2019 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Event*, 2019.

[5] L. Cances, T. Pellegrini and P. Guyot, “Multi-task learning and post processing optimization for sound event detection,” *IEEE AASP Challenge on DCASE 2019 Technical Report*, 2019.

[6] G. Huang, Z. Liu, L. Maaten and K. Weinberger, “Densely connected convolutional networks,” in *Computer Vision and Pattern Recognition 2017*, 2017.

[7] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”, in *Neural Information Processing Systems 2017*, 2017.

[8] R. Serizel, N. Turpault, A. Shah and J. Salamon, “Sound event detection in synthetic domestic environments”, in *International Conference on Acoustics, Speech, and Signal Prcoessing 2020*, Barcelona, Spain, 2020.

[9] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, “Mixup: beyond empirical risk minimization,” in arXiv:1710.09412, 2017.

[10] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Computer Vision and Pattern Recognition 2016*, 2016.

[11] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in arXiv:1412.6980, 2014.

[12] Loshchilov, F. Hutter, “SGDR: stochastic gradient descent with warm restarts,” in arXiv:1608.03983, 2016.

[13] D. Misra, “Mish: a self regularized non-monotonic neural activation function,” in arXiv:1908.08681, 2019.