A CRNN-BASED SYSTEM WITH MIXUP TECHNIQUE FOR LARGE-SCALE WEAKLY LABELED SOUND EVENT DETECTION

Dezhi Wang, Kele Xu, Boqing Zhu, Lilun Zhang, Yuxing Peng, Huaimin Wang*

College of Meteorology and Oceanography, National University of Defense Technology, Changsha, 410073, China
*wang_dezhi@hotmail.com, zll0434@163.com
School of Computer, National University of Defense Technology, Changsha, 410073, China
kelele.xu@gmail.com, zhuboqing09@nudt.edu.cn, pengyuxing@aliyun.com, whm_w@163.com

ABSTRACT

The details of our method submitted to the task 4 of DCASE challenge 2018 are described in this technical report. This task evaluates systems for the detection of sound events in domestic environments using large-scale weakly labeled data. In particular, an architecture based on the framework of convolutional recurrent neural network (CRNN) is utilized to detect the timestamps of all the events in given audio clips where the training audio files have only clip-level labels. In order to take advantage of the large-scale unlabeled in-domain training data, a deep residual network based model (ResNeXt) is first employed to make predictions for weak labels of the unlabeled data. In addition, a mixup technique is applied in model training process, which is believed to have some benefits on the data augmentation and the model generalization capability. Finally, the system achieves 22.05% F1-value in class-wise average metrics for the sound event detection on the provided testing dataset.

***Index Terms*—** DCASE 2018, Weakly-supervised learning, Sound event detection, Convolutional recurrent neural network

1. INTRODUCTION

Great attention has been paid to developing advanced approaches to understand the sounds of everyday life in the contexts of practical applications of smart cars, smart home, surveillance and so on [1-3]. In order to automatically recognize the sound events in an audio recording, sound event detection (SED) has been studied to achieve the temporal information (timestamps) of these events. As the rapid development and significant success of deep learning techniques in recent years, deep learning methods have become the main approaches to solve the SED problem in DCASE challenge [4].

With the purpose of stimulating the development of SED methods in practical applications, Google opens the Audioset (An Ontology And Human-Labeled Dataset For Audio Events) to public, which has a large-scale dataset drawn from Youtube videos [5]. DCASE 2018 Task 4 makes use of a subset of the Audioset to support approaches for the large-scale detection of sound events using weakly labeled data contains only the presence or absence of the audio events (without timestamps). For weakly supervised sound event detection, a large number of deep learning based methods have been developed by using gated convolutional neural networks [6], multiple instance learning [7], sample-level deep convolutional neural networks [8], stacked convolutional and recurrent neural networks [9] and so on. Some architectures proposed by these methods are very complicated and might not be suitable for general applications.

In this report, we explore to develop a simple system based on the CRNN framework using the well-developed neural networks as components to improve the performance and robustness in weakly-supervised sound event detection.

2. PROPOSED METHOD

Generally, we combine several processes together to solve the SED problem, which includes weak label predictions for unlabeled training data, CRNN system for sound event detection, mixup data augmentation and so on.

2.1. Weak label predictions for unlabeled in-domain training data

Since there is only a small weakly annotated training set is provided which is insufficient for an accurate SED in the given context, we utilize a well-developed neural network architecture in the field of computer vision to explore the possibility of making use of a large amount of unbalanced and unlabeled training data to strengthen the system performance. After a lot of comparative studies, the ResNeXt101 model [10] is selected to do the weak label prediction for unlabeled in-domain training data based on the provided weakly annotated training data. As we have little prior knowledge of the out-of-domain unlabeled training data, this part of data is not used in our system. Also we concern that the usage of out-of-domain data could introduce extra noise to the training data especially when the out-of-domain data does not have any reliable annotations.

A 5-fold cross-validation is employed on the weakly labeled training dataset to train the ResNeXt101 model to get a convergence. In the process of weak label prediction, we set a threshold value to keep up to 3 event labels for each clip in the unlabeled in-domain set, which is considered reasonable for the given dataset.

2.2. CRNN-based architecture with multi-time resolution for sound event detection

As shown in Figure 1, after the weak labels are obtained for these unlabeled in-domain training data, we actually obtain an extended weakly labeled training dataset, of which the number of samples has been significantly increased. On the basis of extended training dataset, a CRNN-based system is developed to predict the timestamps of sound events existing in the audio clips of evaluation dataset, which are referred as strong labels. In the CRNN architecture, a well-developed ResNet50 [11] or Xception [12] model is directly used as the CNN component to extract features from the time-frequency representations of input audio data. It should be noted that the ResNet or Xception model has been modified from the original version by reduce the values of stride parameters to 1 in the time axis in the pooling layer (max-pooling or average-pooling). In this way, the pooing operation in the time axis is suppressed and the time-step information is kept as much as possible, which is the basis for timestamp detection.

Following the CNN component is a gated bi-directional RNN layer (using recurrent units, GRUs), where GRU outputs are connected to a gated unit which consists of a sigmoid transform branch and a tanh transform branch. This grated unit is the same one as applied in Google WaveNet [13] model, which is used to introduce the attention mechanism to control the information flow through the networks. Following the gated RNN layer, an additional feed-forward neural network with softmax and sigmoid activations is used to locate the sound events in time axis and finally to achieve the SED output. The configuration of this feed-forward layer can be referred to [6].

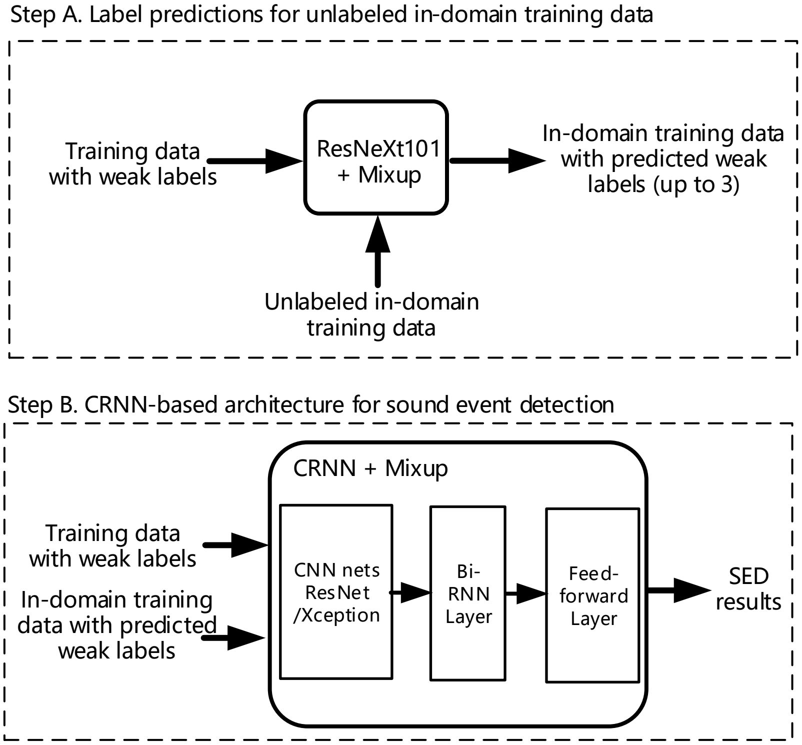
Step A. Label predictions for unlabeled in-domain training data
Training data with weak labels
ResNeXt101 + Mixup
In-domain training data with predicted weak labels (up to 3)
Unlabeled in-domain training data
Step B. CRNN-based architecture for sound event detection
CRNN + Mixup
Training data with weak labels
In-domain training data with predicted weak labels
CNN nets ResNet /Xception
Bi-RNN Layer
Feed-forward Layer
SED results

Figure 1: The two processes of proposed system

2.3. Data augmentation based on mixup

In order to improve the performance of proposed system, an effective data augmentation technique named mixup is applied in both the two processes in our system. Mixup technique theoretically constructs virtual training examples by using the linear combinations of pairs of the representation of examples and their labels [6]. The virtual training examples can be generated by using the following formula:

$$x = \alpha \times x_i + (1-\alpha) \times x_j$$

$$y = \alpha \times y_i + (1-\alpha) \times y_j$$

where (x_i, y_i) and (x_j, y_j) are two samples randomly selected from the training dataset and the α is the mixed ratio normally assigned from the range [0,1]. In our experiment, we choose $\alpha \in Beta(3,3)$. Both the raw wave signals and their time-frequency representations can be used as the mixup samples.

2.4. Data balancing

The data unbalance problem is relatively significant for the data classes in the dataset of DCASE 2018 task 4. We utilize a simple way to do the data balancing just by ensuring at least one sample to be selected for each class in a batch.

2.5. Threshold adjustment

A threshold is needed in order to decide the existence of a sound event in an audio clip. It is important to choose a specific threshold for each class in the given task for SED. We manually adjust the thresholds for each class on the testing dataset. However, it may result in over-fitting problem and has an effect on the performance of trained model on evaluation dataset.

2.6. Model ensemble

It is important to improve the system robustness by doing the fusion of different system results. In this work, we utilize the mean probability strategy to ensemble the probability outputs of the systems with different configurations.

3. EXPERIMENTS

3.1. Datasets and pre-processing

The dataset for 2018 DCASE task 4 is obtained from the Google Audioset, which contains a training dataset (containing 1578 weakly labeled clips, 14412 unlabeled in-domain clips and 39999 unlabeled out-of-domain clips), a testing dataset of 288 clips with strong labels and an evaluation dataset of 880 clips to be predicted. We first down-sample all these audio clips from 44.1 kHz to 16 kHz and transform the wave forms into log-mel energies. Then the 1st and 2nd order delta features of the log-mel features are obtained and all these 3 parts of features are stacked together as a 3-channel feature representation for the audios. Finally, the 3-channel features are used as the input of the proposed system.

3.2. Experimental setup

Different configurations of the proposed architecture are tested in this work. To optimize the loss, the parameters of all the networks are generally tuned depending on the heuristic experience. The label prediction process is carried out based on a 5-fold cross-validation setup and the threshold for label selection is set to 0.3. In the SED process, both ResNet50 and Xception models are used as the CNN component in CRNN framework. The two models are modified to generate time-frequency outputs with different sizes in the time axis. The sizes are set as 120, 128, 240 and 256. Thus, the four submissions are corresponding to the different time resolutions. There is no cross-validation for SED process, only the testing dataset with strong labels is used for test.

4. RESULTS

The SED results of the proposed system on the development testing dataset are listed in Table 1. The F-value and error rate measures are evaluated by using the official sed_val package [14] with a 200ms collar on onsets and a 200ms / 20% of the events length collar on offsets. The four submissions are respectively the fusions of system results where the difference among submissions is located in the time resolutions in sound event detection.

5. CONCLUSIONS

In this work, we have investigated the use of a CRNN-based system integrated with well-developed CNN models and mixup technique for the task 4 of DCASE2018 challenge. The proposed system finally achieve a F1-value as 22.05% which is significantly better than the baseline system which obtain F1 as 14.06%. In future we plan to test the system on other publicly available datasets and continuously make the improvements.

6. ACKNOWLEDGMENT

This work was funded by the Scientific Research Project of NUDT (No.ZK17-03-31) and the National Key R&D Program of China (No.2016YFC1401800).

Table 1: The class-wise metrics including macro-average F1-measure and error rate for Task 4

	Submission 1 using 120 time intervals		**Submission 2 using 128 time intervals**		**Submission 3 using 256 time intervals**		**Submission 4 using 240 time intervals**	
Class	F1-score	Error rate	F1-score	Error rate	F1-score	Error rate	F1-score	Error rate
Alarm/bell	**6.0%**	1.40	**6.0%**	1.41	4.7%	1.45	4.7%	1.46
Blender	**33.8%**	1.10	**33.8%**	1.10	29.9%	1.21	30.3%	1.18
Cat	0.0%	1.87	0.0%	1.87	**1.1%**	1.95	**1.1%**	1.94
Dishes	0.0%	1.43	0.0%	1.55	**3.4%**	1.47	2.3%	1.49
Dog	0.0%	1.68	**0.9%**	1.75	0.8%	1.84	**0.9%**	1.80
Electric shaver	**57.8%**	0.76	**57.8%**	0.76	**57.8%**	0.76	**57.8%**	0.76
Frying	**46.7%**	1.33	45.2%	1.42	38.1%	1.62	38.1%	1.62
Running water	**14.5%**	1.81	13.9%	1.89	13.4%	1.97	13.4%	1.97
Speech	3.4%	1.38	4.4%	1.38	**5.3%**	1.42	**5.3%**	1.43
Vacuum cleaner	**58.4%**	1.06	57.8%	1.09	49.0%	1.43	47.4%	1.46
Average	**22.05%**	1.38	21.98%	1.42	20.35%	1.51	20.12%	1.51

7. REFERENCES

[1] Mesaros, A., T. Heittola, and T. Virtanen. *TUT database for acoustic scene classification and sound event detection*. in *Signal Processing Conference*. 2016.

[2] Ntalampiras, S., I. Potamitis, and N. Fakotakis, On acoustic surveillance of hazardous situations. Acoustics, Speech and Signal Processing, ICASSP 2009: p. 165-168.

[3] Clavel, C., T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. in IEEE International Conference on Multimedia and Expo. 2005.

[4] Mesaros, A., et al., DCASE 2017 challenge setup: Tasks, datasets, and baseline system. Detection and Classification of Acoustic Scenes and Events, 2017.

[5] Gemmeke, J.F., et al. Audio Set: An ontology and human-labeled dataset for audio events. in IEEE International Conference on Acoustics, Speech and Signal Processing. 2017.

[6] Xu, Y., et al., Large-scale weakly supervised audio classification using gated convolutional neural network. arXiv.org, 2017.

[7] Salamon, J., et al., DCASE 2017 Submission: Multiple Instance Learning For Sound Event Detection. Detection and Classification of Acoustic Scenes and Events, 2017.

[8] Lee, J., et al., Combining Multi-Scale Features Using Sample-Level Deep Convolutional Neural Networks For Weakly

Supervised Sound Event Detection. Detection and Classification of Acoustic Scenes and Events, 2017.

[9] Adavanne, S. and T. Virtanen, Sound Event Detection Using Weakly Labeled Dataset With Stacked Convolutional And Recurrent Neural Network. Detection and Classification of Acoustic Scenes and Events, 2017.

[10] Hitawala, S., Evaluating ResNeXt Model Architecture for Image Classification. arXiv.org, 2018.

[11] He, K., et al. Deep Residual Learning for Image Recognition. in IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. in IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[13] Paine, T., et al., Fast Wavenet Generation Algorithm. arXiv.org, 2017.

[14] Mesaros, A., T. Heittola, and T. Virtanen, Metrics for polyphonic sound event detection. Applied Sciences, 2016.