Sound Event Detection Using Weakly Labeled Dataset with Convolutional Recurrent Neural Network

Technical Report

Avdeeva Anastasia

ITMO University
SIS Dept., 49, Kronverkskiy pr., Saint-Petersburg, 197101, Russian Federation
avdeeva-a@speechpro.com

Agafonov Iurii

ITMO University
SIS Dept., 49, Kronverkskiy pr., Saint-Petersburg, 197101, Russian Federation
agafonov@speechpro.com

ABSTRACT

In this paper, a sound event detection system is proposed. This system uses fusion of CNN classifier and CRNN segmentator.

Index Terms — sound event detection, Convolutional Neural Networks, Recurrent Neural Networks.

1. INTRODUCTION

Acoustic event detection is an application of pattern recognition and machine learning in which an audio signal is mapped to corresponding sound events present in the auditory scene. Among others, Sound Event Detection (SED) is a particularly challenging task because it predicts not only possible events but also their start and end times. This task is also particularly challenging because it involves multi-label classification.

Recent deep learning based SED systems that use timestamps information can be divided into two approaches. One is using the sequence information to predict the order of the timestamps, for example, using Recurrent Neural Networks (RNN) [1]. The other is dividing an audio clip into the same length of small segments (e.g. 1 second long) and using the segments as input for the models, for example, using Deep Neural Networks (DNN) or Convolutional Neural Networks (CNN).

In this contribution, a system is proposed which uses fusion of CNN classifier for events class prediction and CRNN segmentator for their start and end time prediction. Actually, we propose two systems which differ only in post-processing stage.

2. SYSTEM OVERVIEW

This section explains approach to classification audio used in this system.

General propositions. A simple activity detection algorithm was applied to all audio data to cut silence. Then all the samples that are below the selected threshold are discarded.

All audio files were resampled to 16 KHz. Log-melspectograms were used as features with the following parameters: STFT window size 512, hop length 256, number of mels were 64 for classifier and 40 for segmentator , time duration 10 seconds. If audio signal was shorter than 10 seconds but larger than 8 it was padded with minimal value.

Augmentation. Because the training dataset is not balanced by classes, and also has a small enough size, two types of augmentation were applied to the data: time stretch and pitch shifting [2].

Architectures. As was mentioned above proposed system is fusion of CNN and CRNN models. The architectures of both models are described below.

CRNN:

Layer
Input: Log mel-band energy (625x40)
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x4
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x4
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x2
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x2
128 units, Bidirectional GRU, tanh
128 units, Bidirectional GRU, tanh
Conv2D: 10 filters, 1x64, Sigmoid

CNN:

Layer
Input: Log mel-band energy (625x64)
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x4
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x4
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x2
Conv2D: 64 filters, 3x3, ReLU
MaxPool2D: 1x2
Conv2D: 10 filters, 1x64, Sigmoid
AveragePooling()

Training. CRNN was trained on a part of labeled training set consist of clips where only one class presence. To get the segmentation mask energy based activity detector was applied to such clips. One exception was clips contain speech. As far as there are no examples of only speech presence in the training dataset, for such examples mask is the same for all classes presence in the clip. Similarly mask was received from activity detector.

CNN was trained on labeled training set to predict classes contains in the clip.

Fusion. Fusion was organized in following way. CRNN-segmentator predicts mask size of 10x625, where the lines represent a markup for each 10 classes. The mask is binarised through thresholds, chosen by applying validation set. CNN-classifier predicts classes that are present in audio recording and based on this prediction the classes in the mask, which are not in the classifier's prediction, reset.

Post-processing. In post-processing stage median filter with window size approximately equal 150ms was applied to the mask received from CRNN.

Also we noticed that some classes such as Blender, Electric shaver toothbrush and Vacuum cleaner can be considered as stationary in comparison with the rest. So we came up with an idea to use for segmentation clips contain these classes spectral flow filter based on autocorrelation. Filters parameters were adopted on validation set. Performance of system with applying such filter for stationary signals is presented in Table 2. The results show that we get slightly better score for some classes.

3. RESULTS

The experimental results (f1-score) are summarized in Table 1 and Table 2 below. These results were received on validation part of development dataset.

Table 1. F1-score for first submission.

Overall f1-score: 26.10%	
Alarm bell ringing	17.8%
Blender	23.4%
Cat	29.5%
Dishes	8.6%
Dog	25.2%
Electric shaver toothbrush	18.5%
Frying	28.6%
Running water	31.9%
Speech	23.5%
Vacuum cleaner	60.9%

Table 2. F1-score for second submission.

Overall f1-score: 28.13%	
Alarm bell ringing	17.8%
Blender	22.0%
Cat	29.5%
Dishes	8.6%
Dog	25.2%
Electric shaver toothbrush	38.7%
Frying	28.6%
Running water	31.9%
Speech	23.5%
Vacuum cleaner	55.6%

4. EVALUATION

Results are evaluated with event-based measures with a 200ms collar on onsets and a 200ms / 20% of the events length collar on offsets. [3]

5. REFERENCES

[1] S. Adavanne, G. Parascandolo, P. Pertil, T. Heittola, and T. Virtanen, “Sound event detection in multichannel audio using spatial and harmonic features,” DCASE2016 Challenge, Tech. Rep., September 2016.

[2] J. Salamon, J. P. BelloDeep, “Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification”, IEEE SIGNAL PROCESSING LETTERS, 2016.

[3] A. Mesaros, T. Heittola, T. Virtanen, Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016.