SURREY-CVSSP SYSTEM FOR DCASE2017 CHALLENGE TASK4

Yong Xu[*], *Qiuqiang Kong*[*]

University of Surrey
Centre for Vision, Speech and Signal Processing
Guildford, GU2 7XH, UK
{q.kong, yong.xu}@surrey.ac.uk

Wenwu Wang, Mark D. Plumbley

University of Surrey
Centre for Vision, Speech and Signal Processing
Guildford, GU2 7XH, UK
{w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

In this technique report, we present a bunch of methods for the task 4 of Detection and Classification of Acoustic Scenes and Events 2017 (DCASE2017) challenge. This task evaluates systems for the large-scale detection of sound events using weakly labeled training data. The data are YouTube video excerpts focusing on transportation and warnings due to their industry applications. There are two tasks, audio tagging and sound event detection from weakly labeled data. Convolutional neural network (CNN) and gated recurrent unit (GRU) based recurrent neural network (RNN) are adopted as our basic framework. We proposed a learnable gating activation function for selecting informative local features. Attention-based scheme is used for localizing the specific events in a weakly-supervised mode. A new batch-level balancing strategy is also proposed to tackle the data unbalancing problem. Fusion of posteriors from different systems are found effective to improve the performance. In a summary, we get 61% F-value for the audio tagging subtask and 0.73 error rate (ER) for the sound event detection subtask on the development set. While the official multilayer perceptron (MLP) based baseline just obtained 13.1% F-value for the audio tagging and 1.02 for the sound event detection. Finally, we ranked first in the audio tagging sub-task on the evaluation set. We also ranked 2nd as a team in the sound event detection sub-task.

Index Terms— DCASE2017, convolutional neural network, attention, audio tagging, sound event detection, weakly labelled data

1. INTRODUCTION

DCASE challenge has been organized for three rounds since 2013 [1, 2, 3]. It urges lots of academic and industrial effort to make analysis and utilizing of the environmental audio. This series of challenges consists three parts, including acoustic scene classification [2], sound event detection [2] and audio tagging [4, 5].

This technique report summaries the methods we use on the task 4 of DCASE2017, the large-scale weakly supervised sound event detection. Task 4 dataset of DCASE 2017 is a subset of Google Audioset [6], consisting recordings of a variety of cars. Task 4 consists of two subtasks, audio tagging and sound event detection. Both of the two subtasks are based on weakly labelled audio data. For audio tagging, lots of deep learning based methods have been developed [7, 8, 9, 10]. As for weakly supervised sound event

detection, multiple instance learning [11], joint detection and classification (JDC) [12] and attention-based methods [13] have been proposed. These weakly labelled learning methods such as the JDC and attention methods have not been well evaluated before the task 4 of DCASE 2017 appear.

In this report, we will present our methods both for audio tagging and weakly-supervised sound event detection.

2. PROPOSED METHODS FOR AUDIO TAGGING

We present a bunch of methods to solve this problem, including data balancing, gated activation function, auto threshold and posterior fusion.

2.1. Data balancing

There are 17 classes in the whole dataset[1]. However the number of samples of each class is not balanced. For example, the sample number of "car" is 25744 while "Car alarm" only has 273 samples. Note that the neural network model is trained in mini batch with stochastic gradient decent. The batch size is always fixed, for example, 128. The 128 samples will be randomly selected from the whole dataset, which means the class with less samples will have low chance to be selected. The neural network is easily biased by the audio samples with large number. Hence, we redesign the sample selection scheme for generating each batch. We almost follow the distribution ratio of each class in the whole dataset, but we also make sure that there is at least one sample for the low probability event in each batch.

2.2. Learnable gated activation function

Sigmoid or ReLU might be used for the activation function in the neural network to introduce the non-linear characteristics. However, in this work, we use a learnable gated activation function [14].

$$\mathbf{Y1} = linear(\mathbf{X} * \mathbf{W} + b) \tag{1}$$

$$\mathbf{Y2} = sigmoid(\mathbf{X} * \mathbf{V} + c) \tag{2}$$

$$\mathbf{Y} = \mathbf{Y1} \otimes \mathbf{Y2} \tag{3}$$

Where $\mathbf{X}$ is the input feature, $\mathbf{Y1}$ is the linear output with the parameters $\mathbf{W1}$. $\mathbf{Y2}$ is the learned gating matrix through parameters $\mathbf{W2}$. Finally, the output $\mathbf{Y}$ is obtained by the element-wise multiplication of $\mathbf{Y1}$ and $\mathbf{Y2}$. This gating activation function can be regarded as a local attention scheme.

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-large-scale-sound-event-detection

2.3. Auto threshold

To decide the existence or presence of an acoustic event, we should choose a hard threshold. Basically, a fixed value will be chosen for each class. However, each class should have a specific threshold. We tuned these thresholds for each class on the development set, and applied them onto the private set.

2.4. System fusion

Fusion is always important to get a robust result. We use two-level fusion. One is to fuse the iterations among the same model. The 2nd level is to fuse the posteriors of different models.

2.5. Whole framework

We treat one CNN layer followed by one batch normalization layer and the learned gating layer as a block. Two this kind of blocks and one max-pooling layer are a unit. Four units were concatenated together to form the whole system. The final output layer is the weighted mean along the time axis to get the 17 output nodes.

3. PROPOSED METHODS FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

The basic framework is similar to the system for the audio tagging. However, we use an additional bidirectional GRU-RNN after the CNN layers. We also did not execute any pooling operations through the time axis to keep the location information.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

The task employs a subset of Google Audioset [6]. AudioSet consists of an expanding ontology of 632 sound event classes and a collection of 2 million human-labeled 10-second sound clips (less than 21% are shorter than 10-seconds) drawn from 2 million YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds. The subset consists of 17 sound events divided into two categories: "Warning" and "Vehicle". The list below shows each class and next to it the approximate number of 10-second clips. Note that each clip may correspond to more than one sound event.

We extracted spectrogram, log-mel filter banks, Mel-frequency cepstrum coefficients (MFCC). Each utterance has 240 frames.

4.2. Results

In this section, the audio tagging results and the weakly supervised sound event detection results will be given.

4.2.1. Audio tagging

Table 1 presents the F1, Precision and Recall comparisons for the **audio tagging** sub-task on the development set and the evaluation set. We finally ranked first in the audio tagging subtasks among 29 systems.

Table 1: F1, Precision and Recall comparisons for the **audio tagging** sub-task on the development the evaluation sets.

Dev-set	F1	Precision	Recall
DCASE2017 Baseline [?]	10.9	7.8	17.5
CRNN-logMel-noBatchBal	42.0	47.1	38.0
CRNN-logMel	52.8	49.9	56.1
Gated-CRNN-logMel (i)	56.7	53.8	**60.1**
Gated-CRNN-MFCC (ii)	52.1	51.7	52.5
Fusion (i+ii)	**57.7**	**56.5**	58.9
Eval-set	**F1**	**Precision**	**Recall**
DCASE2017 Baseline	18.2	15.0	23.1
CNN-ensemble	52.6	**69.7**	42.3
Frame-CNN	49.0	53.8	45.0
Our gated-CRNN-logMel	54.2	58.9	50.2
Our fusion system	**55.6**	61.4	**50.8**

4.2.2. Weakly supervised sound event detection (SED)

The results of F1 and Error rate comparisons on the development set and the evaluation set for the 2nd SED task are given in Table 2. With the fusion system, we rank 2nd as a team in the sound event detection sub-task.

Table 2: The results of F1 and Error rate comparisons on the development set and the evaluation set for the **sound event detection** sub-task among several methods across the 17 audio event tags.

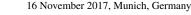| Dev-set	F1	Error rate
DCASE2017 baseline	13.8	1.02
Gated-CRNN-logMel	47.20	0.76
Fusion	**49.7**	**0.72**
Eval-set	**F1**	**Error rate**
DCASE2017 baseline	28.4	0.93
Gated-CRNN-logMel	47.50	0.78
Fusion	**51.8**	**0.73**

Fig. 1 shows an example for predicting locations along 240 frames for "10i60V1RZkQ_210.000_220.000.wav" using the proposed attention method [13]. The above figure is the predictions and the bottom figure is the reference. It shows that the proposed method can predict the locations along the time axis.

5. CONCLUSIONS

This technique report briefly presents the whole system and methods for the task 4 of DCASE2017 challenge. It can get 61% F-value for the audio tagging and 0.73 ER for the weakly supervised sound event detection subtask on the development set. Finally, we ranked first in the audio tagging sub-task on the evaluation set. We also ranked 2nd as a team in the sound event detection sub-task. For more details, please read our paper [15].

6. REFERENCES

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an IEEE AASP challenge," in
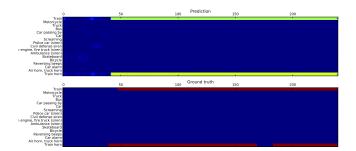
Figure 1: An example for predicting locations along 240 frames for "10i60V1RZkQ_210.000_220.000.wav" using the proposed attention method [13]. The above figure is the predictions and the bottom figure is the reference.

2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4.

[2] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[4] http://www.cs.tut.fi/sgn/arg/dcase2016/task-audio-tagging.

[5] P. Foster and T. Heittola, "DCASE2016 baseline system," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016) challenge*. [Online]. Available: https://github.com/pafoster/dcase2016_task4/tree/master/baseline

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of ICASSP*, 2017.

[7] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *2017 IEEE International Joint Conference on Neural Networks (IJCNN 2017)*.

[8] Y. Xu, Q. Huang, W. Wang, , P. Foster, S. Sigtia, P. Jackson, and M. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," in *IEEE/ACM Trans. on audio, speech and language processing*, 2017.

[9] Y. Xu, Q. Huang, W. Wang, P. Jackson, and M. Plumbley, "Fully dnn-based multi-label regression for audio tagging," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 110–114.

[10] E. Cakir, T. Heittola, and T. Virtanen, "Domestic audio tagging with convolutional neural networks," *DCASE2016 Challenge*, Tech. Rep., 2016. [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2016/documents/challenge_technical_reports/Task4/Cakir_2016_task4.pdf

[11] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[12] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," *Proceedings of ICASSP*, 2017.

[13] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in *Interspeech2017*.

[14] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[15] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.