USTC-NELSLIP SYSTEM FOR DCASE 2018 CHALLENGE TASK 4

Technical Report

Yaming Liu, Jie Yan**

University of Science and Technology of China
National Engineering Laboratory for
Speech and Language Information Processing
Hefei, China
{lym66, yanjie17}@mail.ustc.edu.cn

Yan Song, Jun Du

University of Science and Technology of China
National Engineering Laboratory for
Speech and Language Information Processing
Hefei, China
{songy, jundu}@ustc.edu.cn

ABSTRACT

In this technical report, we present a group of methods for the task 4 of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE 2018). This task aims to detect sound events in domestic environments using weakly labeled training data in a semi-supervised way. In this report, firstly, an event activity detection technique is performed to transform weak labels to strong labels before training. Then a capsule based method and a gated convolutional neural networks (CNN) are performed to estimate event activity probabilities respectively. At last, event activity probabilities of two systems are combined to obtain the final sound event detection (SED) estimation. On the other hand, a tagging model based on proposed CNN is used to tag the unlabeled in domain training set. Data with high confidence are added to the training data to get a further promotion of performance. Experiments on the validation dataset show that the proposed approach obtains an F1-score of 51.60% and an error rate of 0.93, outperforming the baseline of 14.06% and 1.54.

Index Terms— DCASE 2018, Sound event detection, weak label, semi-supervised, Capsule-RNN, convolutional neural networks, audio tagging

1. INTRODUCTION

DCASE challenge has been organized for four rounds since 2013 [1, 2, 3, 4]. Generally, the challenge can be divided into three topic, including acoustic scene classification (ASC), audio tagging (AT) and sound event detection (SED). Task 4 [5] of DCASE 2018 is a SED task which aims to detect sound events of domestic environments using few weakly labeled training data and many unlabel data. Dataset of task 4 are a subset of AudioSet [6]. The latter is collected from YouTube videos uploaded by different users. [1]

2. PROPOSED METHODS

2.1. Proposed capsule based neural network

The capsule based neural network use the same structure with [7]. Firstly, a stack of convolutional layers are designed to extract local features from the input log mel band energies. Then the outputs of CNN are fed into two capsule layers, where local features from different frequency bands and channels are selected to predict multiple objects. A RNN is further applied to model the temporal dependency of capsule layers's outputs. At last, capsule layers and recurrent layers are jointly trained with two different loss functions concurrently to learn effective capsule representation.

2.2. Proposed convolutional neural network

The proposed convolutional neural network is illustrated in Fig. 1. In this network, we consider one convolution layer with gated activation function as a block. This network is similar with [8]. The CNN part which is used as feature extractor consists of four units, and each unit contains several blocks. Considering that semantic features of different sound event have different time duration, we use three windows with different size to partition the output CNN layers. Afterwards, these windows are fed into bi-directional gated recurrent units (GRU). There are two outputs in this network. The output from GRU followed by dense layers with sigmoid activation is considered as sound event detection result. This output can be used to predict event activity probabilities. The other output is the weighted mean of previous mentioned output, considering as audio tagging result. Final loss of the network is the weighted sum of these two outputs. Binary cross-entropy is used for both two outputs.

2.3. Event activity detection

An event activity detection (EAD) technique based energy is used to strengthen the weak label. Specifically, energy of each frame e_t is calculated based log mel band energies. Then a threshold th_k is used to determine whether the frame belongs to event k. Threshold th_k is determined by multiplying the average of energies $\overline{e}$ by a coefficient α_k. T is number of frames of a clip, which is 500 in our experiments.

$$\overline{e}_k = \frac{1}{T}\sum_{t=1}^{T} e_t^k \tag{1}$$

$$th_k = \alpha_k * \overline{e}_k \tag{2}$$

We set α to 1.2 for "foreground" events like *Dog* and 0.6 for "background" events like *Vacuum_cleaner* empirically. In our experiments, events *Cat, Dishes, Dog, Speech* are consider as "foreground" events, and others as "background" events.

[1] The first two authors have equal contribution.

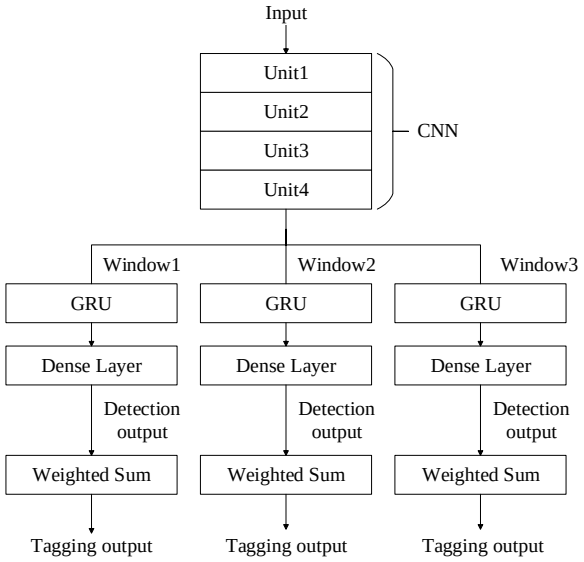

Figure 1: Architecture of the proposed gated convolutional neural network.

For the single-event clips (984 in 1578 clips, only contain one event inside the clip), EAD is computed along whole frequency bands. For the *Speech - Running_water* clips (96 in 1578 clips, contain *Speech* and *Running_water only*), EAD is calculated on the low frequency bands for *Speech* and on the high frequency bands for *Running_water*, respectively. Clips which contain other multi-events are not used in the training process. At last, a post process very similar to section 2.6 is performed to smooth the EAD results.

2.4. Semi-supervised training

A audio tagging model using proposed gated convolutional neural network is trained to tag the unlabel in domain training data. Threshold of tagging model is set to 0.9 to get weakly labeled data with high confidence. These labeled data are further added to the training data.

2.5. Fusion

Two kinds of fusion methods are used in our experiments. Firstly, models of different epoch under same approach are fused to obtain steadier results. Secondly, models of diverse approaches are fused to get the final posteriors of detection.

2.6. Post process

A post process method is used to smooth the posteriors to obtain better event-based F1-score. Specifically, we allow a minimum interval of n frames between two events and a minimum length of m frames of an event. Minimum length m is different for each event and is selected on validation set by grid search, so is minimum interval n.

Table 1: Event-based F1-score of each class (%).

Event	$CRNN$	$CapsRNN$	$Fusion$
Alarm bell ringing	40.4	26.2	40.4
Blender	48.6	29.5	48.7
Cat	50.3	55.8	61.6
Dishes	22.7	20.9	25.7
Dog	53.4	41.9	53.4
Electric shaver toothbrush	55.6	52.0	58.2
Frying	49.1	45.0	51.9
Running water	46.4	52.2	54.4
Speech	33.6	45.8	48.1
Vacuum clearner	73.5	54.8	73.7
Overall	47.4	42.4	51.6

2.7. Dynamic threshold

Threshold is important to the performance of SED system since number of events in each frame is unknown [7]. In our experiments, an optimal threshold is chosen for each event on the validation set.

2.8. Data balance

Data balance technique in [8] is used to relieve the unbalance in training data.

3. EXPERIMENTS AND RESULTS

3.1. Experiments setup

Dataset of task 4 is a subset of Audioset [6], and can be divided into three parts, including training data, validation data and test data. The training data include weakly labeled training data (1578 clips), unlabeled in domain training data (14412 clips) and unlabeled out of domain training data (39999 clips). Length of each clip is equal or less to 10 seconds (less than 21% are shorter than 10 seconds). The subset consists of 10 sound events in domestic environments. Each clip may corresponds to one or more events, so several events may overlapped in the same clip.

In our experiments, 80 bands log mel energies with 40ms frame length and 20ms overlap are calculated as features. Afterwards, each mel band is normalized by subtracting its mean and dividing by its standard deviation calculated over the training set.

3.2. Results

Table 1 shows the event-based F1-score of each class in different models. CRNN is the proposed convolutional neural network. CapsRNN is the proposed capsule based neural network. The third column is the fusion of CRNN and CapsRNN.

4. CONCLUSIONS

The technique report present a system for DCASE 2018 task 4. In our system, an event activity detection technique is performed to transform weak labels to strong labels. Two different approaches are designed to estimate event activity probabilities respectively. Finally we combine these two approaches by fusing their posteriors.

5. 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 *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, 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] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[4] http://dcase.community/challenge2018/.

[5] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," 2018.

[6] J. F. Gemmeke, D. P. 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[7] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule based approach for polyphonic sound event detection," *arXiv preprint arXiv:1807.07436*, 2018.

[8] 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.