CLASS WISE FUSION SYSTEM FOR DCASE 2019 TASK4

Technical Report

Bolun Wang[1]*, Hao Wu*[1]*, Jisheng Bai*[2]*, Chen Chen*[2]*, Mou Wang*[3]*, Rui Wang*[3]
Zhonghua Fu[1]*, Jianfeng Chen*[2]*, Susanto Rahardja*[3] *Xiaolei Zhang*[3]

{blwang, 1070232245, baijs, cc_chen524, wangmou21, wangrui2018}@mail.nwpu.edu.cn
{mailfzh, cjf, susanto, xiaolei.zhang}@nwpu.edu.cn

ABSTRACT

In this report, we introduce our system for Task4 of Dcase 2019 challenges(Sound event detection in domestic environments). The goal of the task is to evaluate systems for the detection of sound events using real data either weakly labeled or unlabeled, along with strong labeled simulated data. With the aim of improving performance with large amount of unlabeled data, and a small labeled training data. We focus on three parts: data augmentation, loss function, and network fusion.

Index Terms— sound event detection, CNN, data augmentation

1. INTRODUCTION

Dcase is a challenge on Detection and Classification of Acoustic Scenes and Events for several years[1]. Dcase 2019 challenge[2] aims at detecting event with small labeled data either weakly labeled or strong labeled, together with a large unlabelled data from Audioset[3] and synthetic recordings. In this paper, we make some exploration for task4 in DCASE 2019 challenge. Firstly we use some data augmentation methods such as event addition, and tagging augmentation, to enrich the data set. Secondly We use log-mel spectrogram as input feature, then test on several network architectures, including CNN, CRNN. Instead of using traditional cross entropy loss function in multi-label classification tasks, we try to optimize the F1 directly. Finally, we find that different classes have different result. There are no such thing as one general model to perform perfectly for every class, so we make a fusion to get a better performance. The paper is organized as follows: In section 2, we make some description for the original data set. In section 3, we show some methods of our system in detail, including data augmentation methods, network models, loss function modification methods, and fusion methods. In section 4, we show the experiment result and make some discussion in case of further research on related topics.

2. DATASET

2.1. dataset of DCASE 2019

The data set of task4 consists of tree parts: weakly labeled data, strong labeled data and unlabeled data. The target of task4 is 10 classes in domestic environment.

Weakly labeled data set contains 1578 clips, with 2244 class occurrences. The unlabeled data contains 14412 clips, The clips are selected such that the distribution per class is close to the distribution in the labeled set. The strong labeled data consists of 2045 clips, with 6032 class occurrences.

2.2. audio preprocessing

Firstly we resample the audio clips to 32000Hz, extract the log mel-spectrogram feature from the clips by a 64 mel bands, a 1024 hanning windows, 500 hop size, so we get 64 frames per second[4]. The overall pipeline of the system is as follows:

- resample wave files to 32000Hz
- extract events with strong time stamps of synthetic data
- add events randomly into original training data in wave domain
- extract log mel-spectrogram features

3. SYSTEM DETAILS

3.1. event extraction and addition

The data augmentation method has been proven to be very useful in [5], especially to improve the generalization of the neural network. According to this idea, we extract the events segment with the time stamps of strong labelled data, and collect them together as 10 separate classes. after extraction, we get multiple wave segments corresponding to the classes. Before the feature extraction, we add them to the existing training data in time domain, with a random scaling factor ranging from 0 to 1:

$$y = \alpha x + (1 - \alpha)e \tag{1}$$

Where α is the scaling factor, y is the augmented data, x is the original training data, e is the extracted event segment, all in time domain.
And the starting point for addition is a little different. For weakly labeled data, we randomly select a sample point in all samples, then add the event samples from the point, and for strong labeled data, we can add exactly into the time interval with onset time and offset time given.

3.2. audio tagging augmentation

Just like any other data driven methods, more training data leads to higher performance. So we divide our task into two steps.

- audio tagging step
- sound event detection step

Firstly, we use original training weakly labeled data to build a tagging data network, with error rate as lower as possible, this network is then used to do the audio tagging for the large amount unlabeled data.
The audio tagging network is impossible to tag all clips correctly, which may impact the further detection results. Because of this, we set a very high threshold(0.95) so that we may lose some training data, but we have a higher reliability of training data.
Secondly, we feed the original training data and tagged data into the sound event detection network, which is similar to the audio tagging network in structure.

3.3. Neural networks

We tested the performance on multiple networks, including CNN, RNN and the combination of them. our CNN model has 9 layers, similar structure in[4] shows that 9 layers outperform any other configurations.

3.4. F1 loss optimization

The evaluation of task4 is F1 score, instead of the traditional loss function binary cross entropy, the better one would be the f1 it self, then the misalignment disappears[6]. The F1 is not differential, so that we must modify it. Instead of accepting 0/1 integer predictions, we compute the F1 loss on the probability matrix directly, along with binary cross entropy loss function, we define a weighted loss function:

$$J = \beta J_F + (1 - \beta) J_B \tag{2}$$

Where J_B is the binary cross entropy, which is usually used in multiple label classification tasks, and J_F is the proposed F1 loss function. This loss function makes it possible to optimize classification and F1 score simultaneously.

3.5. class wise fusion

There is no such thing as one method works best for all classes, methods leads to different results for different classes, which differs very greatly some times. So we combine multiple result together to reduce prediction error.

4. RESULT AND DISCUSSION

This section shows results of our experiments. Table 1 shows the result of multiple experiments:

- baseline: The official baseline system, which is contributed from the best submission of DCASE 2018 in[7].
- event 1s: the events are extracted to have a fixed length(1 sec), then added to the original wave form during feature extraction.
- event original: the events are extracted by the original length, then event addition is performed with a scaling factor α.
- CRNN: concatenate two RNN layers after the CNN layers
- dynamic threshold: different events have different triggering threshold T, so we tested serveral threshold for audio tagging, default T=0.9, and T=0.8 in Table 1.
- unlabel augmentation: use large amount unlabeled data as back ground noise, which is added to the wave file to improve the generalization performance.

method	event F1	segment F1
baseline	0.237	0.552
event 1s	0.187	0.440
event original	0.247	0.603
CRNN	0.187	0.560
dynamic threshold	0.233	0.591
unlabel augmentation	0.200	0.569
F1 loss function	0.250	0.611
synthetic augmentation	0.223	0.561
fusion	0.319	0.605

Table 1: F1 for different methods

β	event F1	segment F1
1.0	0.208	0.527
0.9	0.250	0.611
0.7	0.214	0.598
0.5	0.240	0.602
0.3	0.220	0.60

Table 2: F1 for different β in loss function experiment

- F1 loss function: change the loss function from original binary cross entropy to new loss function, which combine F1 and binary cross entropy, in Table 1, β=0.9.
- synthetic augmentation: combine synthetic data and weakly data in development set as training data. Instead of using strong label for synthetic data, we use only the weak label.
- fusion: combine all above networks together to improve the performance, with the result on validation set.

Table 2 show the results of different β for the experiment of combination of F1 and original binary cross entropy. The result shows that the optimal β is 0.9.

From the experiment results above, we can get some conclusions:

- fusion method outperform any other single methods, so it's better to combine some methods for better performance.
- different classes have different property, which may lead to totally different result. We can consider only one class every time.

class	event F1
Speech	0.449
Dog	0.145
Cat	0.361
Alarm_bell_ringing	0.188
Dishes	0.132
Frying	0.385
Blender	0.295
Running_water	0.262
Vacuum_cleaner	0.46
Electric_shave_toothbrush	0.516
Overall	0.319

Table 3: Event based F1 for final fusion scheme

- instead of traditional binary cross entropy for multi-label classification tasks, combination of F1 and binary cross entropy get better result.

5. REFERENCES

[1] 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.

[2] http://dcase.community/challenge2019/.

[3] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[4] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[5] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," DCASE2018 Challenge, Tech. Rep., September 2018.

[6] https://www.kaggle.com/rejpalcz/best-loss-function-for-f1-score-metric.

[7] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.