Urban Sound Tagging With Multi-Feature Fusion System

Technical Report

Jisheng Bai[1], *Chen Chen*[1], *Bolun Wang*[2], *Mou Wang*[3], *Rui Wang*[3]

Jianfeng Chen[1], *Zhonghua Fu*[2], *Susanto Rahardja*[3], *Xiaolei Zhang*[3]
{baijs, cc_chen524, blwang, wangmou21 , wangrui2018}@mail.nwpu.edu.cn
{ cjf , mailfzh, susanto, xiaolei.zhang}@nwpu.edu.cn

ABSTRACT

This paper presents a multi-feature fusion system for the DCASE 2019 Task5 Urban Sound Tagging(UST). It focus on predicting whether each of 23 sources of noise pollution is present or absent in a 10-second scene [1]. There are coarse-level and fine -level taxonomies to train model. We mainly focus on coarse-level and use best coarse-level model architecture to train fine-level model. Various features are extracted from original urban sound and Convolutional Neural Networks(CNNs) are applied in this system. Log-Mel, harmonic, short time Fourier transform (STFT) and Mel Frequency Cepstral Coefficents (MFCC) spectrograms are fed into a 5-layer or 9-layer CNN, and a type of gated activation [2] is also used in CNN. Different feature is adapted for different urban sound classification according to the results of our experiment. We get at least 0.14 macro-auprc score improvement compared to baseline system on coarse-level. Finally, we make a fusion of some models and evaluate on evaluation dataset.

Index Terms— DCASE, Urban Sound Tagging, Convolutional Neural Networks, multi-feature fusion

1. INTRODUCTION

The city of New York, like many others, has a "noise code". The noise code presents a plan of legal enforcement and thus mitigation of harmful and disruptive types of sounds. Although harmful levels of noise predominantly affect low-income and unemployed New Yorkers, these residents are the least likely to take the initiative of filing a complaint to the city officials. For reasons of comfort, public health and improving fairness, accountability, and transparency in public policies against noise pollution, to control and learn the distribution of noise is essential for government.

Meanwhile some of the most successful techniques in the challenge could inspire the development of an embedded solution for low-cost and scalable monitoring, analysis, and mitigation of urban noise.

In sound tagging and classification, the CNNs are successful applied and achieve great results such as bird sound detection [3], acoustic scene classification [4] and domestic activities [5] [6]. Log-Mel spectrogram is a common feature and widely used in Detection and Classification of Acoustic Scene and Event [7] [8]. So tagging urban sound based on log-Mel spectrogram and CNNs is supposed to achieve good results.

In our system, firstly, log-Mel, percussive and harmonic, STFT, MFCC spectrograms are extracted as features. Then we experiment different features on a VGG-like network and to recognize the influence between eight coarse classes. After that, a gated activation is further applied for sound event detection. Finally, we evaluate evaluation data and fusion the results referred to the scores between different classes.

This paper is organized as follows: in Section 2, the official datasets and evaluation matrics will be introduced; in Section 3, feature extracted method will be introduced; In Section 4, some network architectures are shown and Section 5 will give the experiment results; conclusion and discussion are followed in Section 6.

And our code is open source on github[1].

2. DATASET AND EVALUATION METRICS

2.1. Development and evaluation datasets

The development dataset contains a train split (2351 recordings) and validate (443 recordings). These recordings are from SONYC acoustic sensor network for urban noise pollution monitoring. Over 50 different sensors have been deployed in New York City. The train and validate splits are disjoints and it make participants to develop computational systems for multilabel classification in a supervised manner. And validation subset can prevent overfitting during the training.

The reference labels are coarse-level and fine-level taxonomies and each recording is listened at least three humans independently. The relationship of hierarchical containment between coarse-grained and fine-grained taxonomy are shown in figure 1. The evaluation dataset contains 274 recordings and may be from validate split.

2.2. Evaluation matrics

The UST challenge is a task of multilabel classification. The area under the precision-recall curve (AUPRC) is the classification metric to evaluate. And for coarse-grained and fine-grained AUPRC, micro-auprc and macro-auprc are both computed, F-score is used for analysis as well.

[1] https://github.com/CN-BOTK/dcase2019-task5

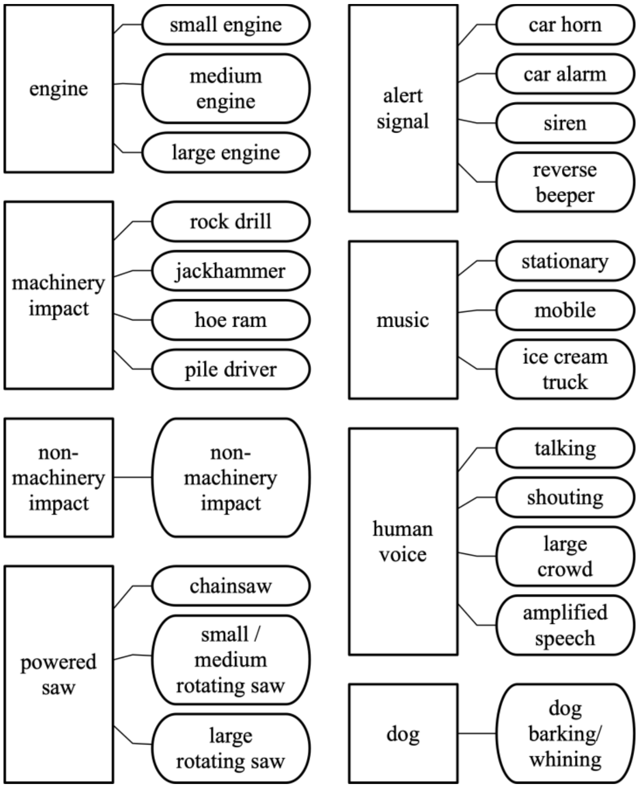

Figure 1: Hierarchical taxonomy of urban sound tags in the DCASE Urban Sound Tagging task. Rectangular and round boxes respectively denote coarse and fine tags.

3. FEATURES

Mainly features are extracted with python librosa functions and described as follows.

3.1. STFT

Recordings fistly are resampled to 16000 Hz .Short time Fourier transform (STFT) spectrograms, using librosa.core.stft with a Hanning window size of 1024 samples and a hop length of 664 samples, are extracted from recordings.

3.2. HPSS

Median-filtering harmonic percussive source separation (HPSS) [9] can separate harmonic and percussive components from input spectrogram (STFT spectrogram). Librosa.decompose.hpss function is applied for feature extracting.

3.3. Log-mel

Recordings are resampled to 32000 Hz and to generate mel spectrogram with a Hanning window size of 1024 and hop length of 500 samples. Mel filters with different bands (64,80 and 128), are used to transformed STFT spectrogram to mel spectrogram, and frequencies lower than 50 Hz and beyond 14000 Hz are removed.

3.4. MFCC

We get MFCC from log-mel spectrogram. And MFCC of 24 n_mfcc and first and second order difference are used as input. All spectrograms are converted to power spectromgrams yielding a dynamic range of 80 dB.

4. UST NETWORK ARCHITECTURE

CNNs have been widely used in computer vision and have achieved state-of-the-art performance in several tasks such as image classifification [11]. The filters can capture local patterns of the input feature maps, such as edges in lower layers and profiles of objects in higher layers. Different network architectures are applied for different features. We experiment CNN, CRNN, inception-v3 and capsnet, and only CNN perform well on development dataset. So we focus on CNN and the architectures are shown in Table 1. The CNN architectures are similar to [10]. Batch normalization [12] is applied to speed up and prevent overfitting during train steps. And leaky_Relu or gated function are used as a non-linear activation after batch normalization. Average pooling with size of 2*2 to reduce the feature map. Then the frequency axis is averaged out and frame axis is maxed out after the last convolutional layer.

For training, Tensorflow is implemented. Sigmoid cross entropy is utilized as loss function and AdamOptimizer as optimizer with a learning rate of 0.001. Training is done with batch size of 32 and we early stop the training if the macro-auprc doesn't improve in last 3 steps.

5. EXPERIMENT RESULTS

5.1. CNN architectures

Log-mel spectrograms with 64 mel bands are fed into different CNN architectures. CRNN3 contains 3 convolutional layers and 2 recurrent layers. The results are shown in Table 2.
Scores in Table 2 shows that the best CNN architecture for UST is CNN9. And CNN9_gated achieve second best result. The CRNN3 is far beyond CRNN9 and approximately same as inception-v3. The worst result is Capsnet.

5.2. Features

As described in Table 1, the best results of every coarse class macro-auprc are shown in Table 3.

5.3. Fusion

In the end, we trained our model with the details shown in Table 1 on the whole development dataset. Evaluation dataset fusion is done by selecting the best model for every coarse class (shown in Table 3). In addition, we trained classes with its best performed features, for example 'engine' and 'powered-saw' are trained with STFT.

Table 1: Feature and network architecture

	CNN5		CNN9	CNN9_gated
Features	STFT	HPSS_h	Log-mel	MFCC
Conv1	3*3@64,BN,Relu		(3*3@64,BN,Relu)*2	(3*3@64,BN,Gated)*2
Pool1	2*2 average pooling			
Conv2	3*3@128,BN,Relu		(3*3@128,BN,Relu)*2	(3*3@128,BN,Gated)*2
Pool2	2*2 average pooling			
Conv3	3*3@256,BN,Relu		(3*3@256,BN,Relu)*2	(3*3@256,BN,Gated)*2
Pool3	2*2 average pooling			
Conv4	3*3@512,BN,Relu		(3*3@512,BN,Relu)*2	(3*3@512,BN,Gated)*2
Pool4	1*1 average pooling			
Dense	512			

Table 2: Coarse -level best performance

	micro-auprc	Micro-f1score	Macro-auprc
Baseline	0.76	0.67	0.54
CNN9	0.82	0.74	0.63
CNN9_gated	0.81	0.72	0.62
CRNN3	0.72	0.66	0.51
Inception-v3	0.72	0.69	0.50
CRNN9	0.51	0.54	0.38
Capsnet	0.54	0.34	0.35

Table 3: class auprc on validate split

Coarse class	Feature	Macro-auprc
1_engine	STFT	0.85
2_machinery	Log-mel	0.54
3_nonmachinery	Log-mel	0.62
4_powered-saw	STFT	0.80
5_alert	Log-mel	0.86
6_music	HPSS_h	0.47
7_human-voice	Log-mel	0.95
8_dog	Mfcc(n_mel=24)	0.33

6. CONCLUSION AND DISCUSSION

By comparing the results in Table 2 and Table 3, some conclusions can be made as follows:

1. CRNN has been proved to be state-of-art method of sound event detection [8], capsnet also achieve best results in sound event detection [13]. Inception-v3 is applied in bird sound detection and improve detection performance. But in UST, these architectures did not work better, even worse. Some reasons could be selecting the large amount of hyper parameters of these architectures, or the defects for UST of them.
2. Different features may be adapt for tagging different source of urban sound. Original STFT spectrogram can explore 'engine' and 'power' feature. Harmonic spectrogram is discovered to recognize 'music' better, because 'music' contains harmonic waves apparently. MFCC can improve 'dog' auprc score. This may inspire us to classificate different sound with unique features rather than one single type.
3. Gated activation [2] can further improve 'dog' macro-auprc compared with leaky_Relu activation in CNN9.
4. As for annotations of development dataset, although there are at least three annotators for every recording, many of the annotations are badly annotated. So the tagging accuracy could be improved if the annotations were more reliable.
5. Further work. Different network architectures and hyper parameter selecting will be studied, and advantages of tagging with different features will be researched as well.

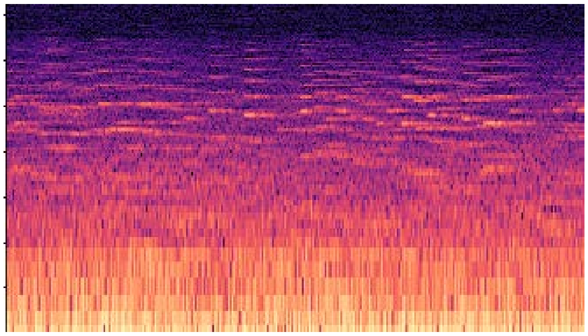

Figure 2: A HPSS_h spectrogram of music.

7. ACKNOWLEDGMENT

I would like to thank Qiuqiang Kong for his primary work of DCASE2019 [10] and New York University for organizing this challenge.

8. REFERENCES

[1] http://dcase.community/challenge2019/task-urban-sound-tagging

[2] Jisheng Bai, et al. CIAIC-BAD SYSTEM FOR DCASE2018 CHALLENGE TASK 3. Tech. Rep.,DCASE2018 Challenge, 2018.

[3] Mario Lasseck. ACOUSTIC BIRD DETECTION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS. Tech. Rep.,DCASE2018 Challenge, 2018.

[4] Yong xu, et al.LARGE-SCALE WEAKLY SUPERVISED AUDIO CLASSIFICATION USING GATED CONVOLUTIONAL NEURAL NETWORK. Tech. Rep.,DCASE2017 Challenge, 2017.

[5] Dexin Li, et al. CIAIC-BAD SYSTEM FOR DCASE2018 CHALLENGE TASK 5. Tech. Rep.,DCASE2018 Challenge, 2018.

[6] Vuegen L., Karsmakers P., Van hamme H., Vanrumste B. (2018). Weakly-Supervised Classification of Domestic Acoustic Events for Indoor Monitoring Applications. Presented at the IEEE Conference on Biomedical and Health Informatics 2018, Las Vegas, Nevada, USA.

[7] Serizel, Romain , et al. "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments." (2018).

[8] Vesperini, Fabio , et al. "Polyphonic Sound Event Detection by using Capsule Neural Networks." (2018).

[9] Fitzgerald, Derry. “Harmonic/percussive separation using median filtering.” 13th International Conference on Digital Audio Effects (DAFX10), Graz, Austria, 2010.

[10] Qiuqiang Kong, et al. Cross-task learning for audio tagging, sound event detection spatial localization: DCASE 2019 baseline systems.(2019)

[11] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” International Conference on Learning Representations (ICLR), 2015.

[12] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internalcovariate shift,” arXiv preprint arXiv:1502.03167, 2015

[13] Liu, Yaming , et al. "A Capsule based Approach for Polyphonic Sound Event Detection." (2018).