A Two-Stage Approach to Device-Robust Acoustic Scene Classification

[Paper No. 2953]

Hu Hu[1], Chao-Han Huck Yang[1], Xianjun Xia[2], Xue Bai[3], Xin Tang[3], Yajian Wang[3], Shutong Niu[3], Li Chai[3], Juanjuan Li[2],
Hongning Zhu[2], Feng Bao[2], Yuanjun Zhao[2], Sabato Marco Siniscalchi[1,4], Yannan Wang[2], Jun Du[3], Chin-Hui Lee[1]
[1]*Georgia Institute of Technology, GA, USA* [2]*Tencent Media Lab, Tencent Corporation, China* [3]*University of Science and Technology of China, HeFei, China* [4]*University of Enna Kore, Italy*

Overview

This work aims to solve the device robustness problem of acoustic scene classification (ASC) tasks. We proposed a novel **Two-Stage ASC system** which leverages on an ad-hoc score combination based on two CNN classifiers. We got a state-of-the-art 81.9% average accuracy among multi-device test data on DCASE 2020 Task 1a development data set.

Keywords: Acoustic scene classification, robustness, convolutional neural networks, data augmentation, class activation mapping.

Two-Stage Procedure

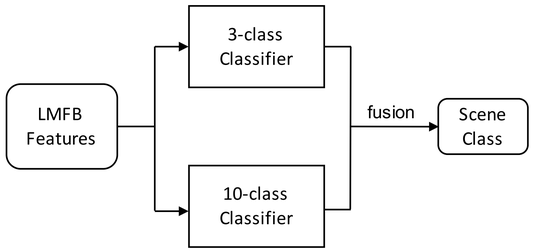


Figure 1: The proposed two-stage ASC system.

Figure 1 shows the proposed two-stage ASC system. The overall system consists of two independent classifiers and outputs the class of the input audio scene choosing among ten classes.

3-class classifier: It classifies an input scene audio into one of three broad classes: in-door, out-door, and transportation. This 3-class classification way is from our prior knowledge that scene audios can be roughly categorized into such three classes.

10-class classifier: The main classifier, which assigns a given input audio clip into one of ten target acoustic scene classes.

The final predicted class $Class(x)$ for the input x is:

$$Class(x) = \underset{q,(p\in\mathbb{C}^1,q\in\mathbb{C}^2,p\supset q)}{\operatorname{argmax}} F_p^1(x) * F_q^2(x),$$

where $p \supset q$ means that p can be thought of a super set of q.

CNN Classifiers

(1) **Resnet**: A duel-path deep residual convolutional neural networks. It has no frequency sub-sampling throughout the whole network. Each input feature map is divided into two sub-feature mapping along the frequency dimension.

(2) **FCNN**: A fully convolutional neural networks built with 9 stacked convolutional layers with small-size kernel. Each convolutional layer is followed by a batch normalization operation and ReLU activation function. Before the final global average pooling layer, channel attention is applied to each output channel of the last layer.

(3) **fsFCNN**: An extension of FCNN model, which mainly has 2 more convolutional layers and reduces max-pooling size in the frequency axis. It can help to leverage over-fitting issue over FCNN model.

Data Augmentation

(1) **Mix-up**.

(2) **Random cropping**: Input time bin is randomly cropped into a smaller fix number.

(3) **SpecAugment**: Input features are randomly masked along time and fequency dimensions.

(4) **Spectrum correction**: Generate new features by a correction coefficient.

(5) **Reverberation with Dynamic Range Compression (DRC)**: RIRs are used to create reverberation, and DRC is then applied to dynamically compress the amplitude range.

(6) **Pitch shift**: Randomly shift the pitch of each audio clip based on the uniform distribution.

(7) **Speed change**: Randomly change the audio speed based on the uniform distribution.

(8) **Random noise**: Add random Gaussian noise.

(9) **Mix audios**: Randomly mix two audios from the same scene class.

Experiments

Data set: DCASE 2020 task1a develop-ment data set. 10-second single-channel train audio clips recorded by 9 different devices.
Acoustic features: 128-dimension log-mel filter bank (LMFB) features, with delta and delta-delta. The utterance-level scaling operation to scale LMFB coefficients into [0,1].
Overall results of proposed systems:

System	A %	B&C %	s1-s3 %	s4-s6 %	Avg. %
Official Baseline	70.6	61.6	53.3	44.3	54.1
Resnet	83.0	76.1	73.6	71.0	74.6
FCNN	87.3	79.5	75.7	73.0	76.9
fsFCNN	83.9	78.6	75.4	72.8	76.2
Ensemble	87.0	81.5	78.0	76.9	79.4
2-stage Resnet	84.5	78.6	76.2	76.4	77.7
2-stage FCNN	89.1	82.9	78.5	76.9	80.1
2-stage fsFCNN	83.9	81.2	78.6	76.4	79.0
2-stage Ensemble	**87.9**	**84.1**	**80.4**	**79.9**	**81.9**

Table 1: Overall results on DCASE 2020 task1a development set.

3-class classification results:

3-class Model	Resnet	FCNN	Ensemble
Acc. %	91.4	92.9	93.2

Table 2: 3-class Classification results

Evaluation results of different data augmentation strategies:

System	A %	B&C %	s1-s3 %	s4-s6 %	Avg. %
Resnet	78.8	72.1	69.3	69.5	71.0
+sa	80.3	73.5	71.4	67.7	71.6
+sa+sc	79.1	75.0	70.7	68.9	72.0
+sa+sc+r	80.3	74.7	71.4	70.3	72.8
+sa+sc+r+aug*	83.0	76.1	73.6	71.0	74.6

Table 3: Accuracy comparison of different data augmentation strategies on Resnet. Mixup and random cropping are always employed. 'sa' indicates specAugment. 'sc' indicates spectrum correction. 'r' indicates reverberation with DRC. 'aug' indicates another four augmentation methods, including pitch shift, speed change, random noise and mix audios.

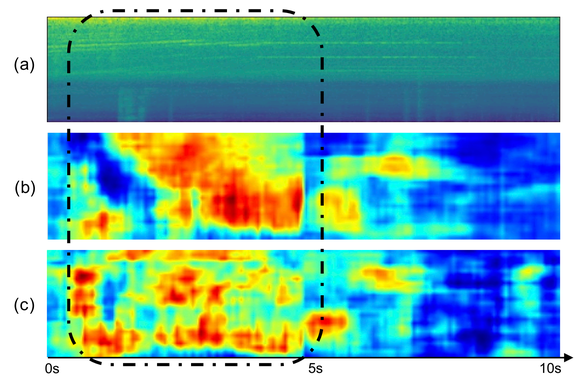
More information

Code available: https://github.com/MihawkHu/DCASE2020_task1

Neural Saliency Analysis

We use CAM to gain a better understanding about what sound patterns are found by our CNNs to accomplish the ASC task.
The results demonstrate that CNNs pay particular attention to audio segments strictly related acoustic events, rather than simply fetching information from the background environmental sound.
The three subplots are: (a) spectrogram, (b) CAM of 3-class classifier, (c) CAM of 10-class classifier

CAM example 1:
Audio: *metro_station_vienna_87_2389_a.wav*
Content: Brake and horn sound starts from 0s to around 8s and only reverberation remains after 5s.



CAM example 2:
Audio: *bus_prague_1102_42431_a.wav*
Content: Brake sound starts from around 2s to around 5s, human talk starts from around 5s to 7s.