Cross-domain sound event detection: from synthesized audio to real audio

Technical Report

Junyong Hao， Zhenwei Hou, Wang Peng

Chongqing University
Chongqing,China
201808021030@cqu.edu.cn

ABSTRACT

This technical report describes some of the system information submitted to dcase2020 task4 - Sound Event Detection in Domestic Environments. We use the dataset of dcase2019 task4 to train our model, contains strongly labeled synthetic data, large unlabeled data, and weakly labeled data. There is a very large domain gaps in the statistical distribution between the synthesized audio and the real audio, and the performance of the SED model obtained on the synthesized audio applied to the real audio is greatly reduced. To perform this task, we propose a DA-CRNN network for joint learning of sound event detection(SED) and domain adaptation(DA).We consider the impact of the distribution within a single sound on the generalization performance of the model by mitigating the impact of complex background noise on event detection and the self-correlation consistency regularization of clip-level sound event classification, these make the intra-domain of a single sound smoother ; for cross-domain adaptation, adversarial learning through feature extraction network with frame-level domain discriminator and clip-level domain discriminator, forcing the feature extraction network to learn the invariant features of the domain, and further improve the generalization performance of the model. We did not use sound source separation, achieved an F1 score of 48.25% on the validation dataset and an F1 score of 49.43% on the public evaluation dataset.

Index Terms— Sound event detection, domain adaptation, consistency regularity

1. INTRODUCTION

Sound event detection (SED) aims to detect and identify each sound event category and its onset and offset in the audio sequence. SED research includes audio event classification, abnormal sound detection, and automatic monitoring. However, there are not many practical applications for sound event detection. Due to the diversity and complexity of real-life sound field environments, sound event detection can only be barely used in a few simple scenarios. The SED task requires a large amount of labeled training data, and these data cost a lot of cost for a large number of people to perform sound event categories and its onset and offset. In order to solve the problem of high cost of acquiring data labels in SED tasks, one solution is to use synthetic audio data to train the model. Current computer technology can synthesize high-quality audio sequences, and can generate a labeled synthetic audio dataset (such as DCASE2019task4) for SED model training. However, because the statistical distribution between the synthesized audio and the real audio has a very large domain gap, the performance of the SED model obtained on the synthesized audio applied to the real audio is greatly reduced.

Domain adaptation relieves the model's overfitting of training set (source domain) data by reducing the difference in statistical distribution across domains, and does not require the assumption that training (source domain) data and test (target domain) data are independently and identically distributed. Using the data of unlabeled real audio, we can understand the difference information of the distribution measurement between the synthesized audio and the real audio. This information can be used to adjust the classification of the model and change the prediction result of the model.

2. PROPOSED METHODS

Convolutional neural network CNN has achieved the best results in the SED task. By transforming the one-dimensional time domain signal of the sound into two-dimensional time-frequency domain signal through FFT, the features of Mel spectrogram are extracted as input to a neural network. Some networks modeled by CNN-RNN have achieved very good recognition results in SED. CNN-RNN uses CNN to extract high-dimensional semantic features of Mel spectrogram, and then these features are sent to RNN for time series modeling, usually there will be tasks after RNN Classifier to complete the classification task.

2.1. Mel spectrogram

The dataset of dcase2019 task 4 is composed of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. We use raw source clips to extract the mel-spectrogram without source separation pre-processing.

First, we resample the audio clips at 22050 Hz, windows is 2048, hop length is 431, number of mels is 64 extraction frequency range between 50hz - 11025hz, the size of the spectrogram is (512,64). Second we noticed that some unlabeled audio files are seriously less than 10s, so we removed audio files <400kb. Finally we normalize the input spectrum to 0-1 through its own maximum and minimum values, rather than the statistics of the entire dataset.

2.2. Training

CRNN uses the Adam optimizer, the initial learning rate is set to 2e-3, and gradually decreases to 1e-5 with training. The two domain discriminator networks also use the Adam optimizer, and the initial learning rate is set to 2e-4, as the training gradually Down to 1e-6; batch size is 8, respectively 2 synthesized audio, 2 weakly labeled audio, 4 unlabeled audio. weight decay is set to 1e^{-5}.

2.3. Evalution and post-processing

We counted the average number of frames of each type of event in the synthesized data, and selected the length of 1/3 as the post-processing median filter window length, combine with audio tagging embedding as post-processing for sound event detection.

We add weak label data in the target domain, adversarial learning by synthesizing audio with unlabeled audio and weakly labeled audio.

dataset(2019)	F1 (%)	Precision(%)	Recall(%)	ER(%)
validation	48.25	53.31	44.72	0.94
public evaluation	49.43	60.45	42.71	0.86

Figure 1: Indicators on different datasets.

dataset(2019)	validation	public evaluation
Alarm	46.1	47.9
Blender	59.5	48.8
Cat	48.9	63.5
Dishes	28.5	35.5
Dog	29.8	40.5
Electric	62.9	55.2
R_water	44.9	35.7
Speech	51.6	59.8
V_cleaner	69.7	57.3
overall	48.25	49.43

Figure 2: Class-wise F1 score(%) on different datasets.

3. REFERENCES

[1] http://dcase.community/workshop2020/.

[2] http://www.ieee.org/web/publications/rights/copyrightmain.html

[3] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, London, UK: Academic Press, 1999.

[4] C. D. Jones, A. B. Smith, and E. F. Roberts, “A sample paper in conference proceedings,” in *Proc. IEEE ICASSP*, 2003, vol. II, pp. 803-806.

[5] A. B. Smith, C. D. Jones, and E. F. Roberts, “A sample paper in journals,” *IEEE Trans. Signal Process.*, vol. 62, pp. 291-294, Jan. 2000.

[6] 2018 Task 4,” Detection and Classification of Acoustics Scenes and Events 2018, Shanghai, China, Jul.2018, pp. 1-5.

[7] Lin L , Wang X , Liu H , et al. Guided Learning Convolution System for DCASE 2019 Task 4[J]. 2019.

[8] Lin L , Wang X , Liu H , et al. Specialized Decision Surface and Disentangled Feature for Weakly-Supervised Polyphonic Sound Event Detection[J]. 2019.

[9] Turpault N , Serizel R , Salamon J , et al. Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis[C]// 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2019). 2019.