A Mobile Application for Sound Event Detection

Yingwei Fu[1,2], **Kele Xu**[1,2*], **Haibo Mi**[1,2], **Huaimin Wang**[1,2], **Dezhi Wang**[3] and **Boqing Zhu**[1,2]
[1]National Key Laboratory of Parallel and Distributed Processing
[2]College of Computer, National University of Defense Technology
[3]College of Meteorology and Oceanography, National University of Defense Technology
yingwei_fu_nudt@163.com, kelele.xu@gmail.com, haibo_mihb@126.com, whm_w@163.com,
wangdezhi08@nudt.edu.cn, zhuboqing09@nudt.edu.cn

Abstract

Sound event detection is intended to analyze and recognize the sound events in audio streams and it has widespread applications in real life. Recently, deep neural networks such as convolutional recurrent neural networks have shown state-of-the-art performance in this task. However, the previous methods were designed and implemented on devices with rich computing resources, and there are few applications on mobile devices. This paper focuses on the solution on the mobile platform for sound event detection. The architecture of the solution includes offline training and online detection. During offline training process, multi model-based distillation method is used to compress model to enable real-time detection. The online detection process includes acquisition of sensor data, processing of audio signals, and detecting and recording of sound events. Finally, we implement an application on the mobile device that can detect sound events in near real time.

1 Introduction

Our living environment contains many types of sound events that provide us with a wealth of useful information to help us identify and perceive the environment [Xu *et al.*, 2018; Zhu *et al.*, 2018; Xu *et al.*, 2019]. The sound event detection (SED) task is proposed to help the intelligent devices understand the sound events and better serve humans. The task of SED includes localization and classification of sound events, aiming at estimating the onset and offset times of sound events and predicting the sound events to predefined types. SED is widely used for many applications. For example, in the field of driverless driving, if the system can identify the sound events of vehicles approaching or leaving, it can make self-driving system more reliable. Except driverless driving, SED is also used for environmental surveillance [Harma *et al.*, 2005] and multimedia events detection [Wang *et al.*, 2016]. In real life, sound events often overlap, which makes it difficult to detect the sound event from a mixture of different sound events. Depending on how to tackle the overlapping sound events in audio, SED tasks can be divided into monophonic sound event detection and polyphonic sound event detection. Monophonic SED only detects the most prominent sound event at a given time while polyphonic SED can detect multiple sound events at the same time which is closer to real-life scenario.

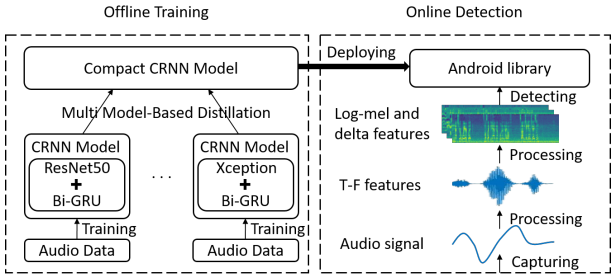

Figure 1: The architecture overview.

Recently, deep neural networks have shown good performance on polyphonic SED task, which use the log mel band energy features or mel frequency cepstral coefficients (MFCC) features as input. Convolutional neural network (CNN) [Espi *et al.*, 2015] can exploit spatially local correlation across input data, and recurrent neural network (RNN) [Parascandolo *et al.*, 2016] can capture long term temporal context for the audio signal. As a result of taking advantage of both approaches, convolutional recurrent neural network (CRNN) has provided state-of-the-art results. However, most of the deep models have millions of parameters, which is not applicable to the embedded or mobile devices with limited computation and storage resources. So there are few applications on mobile devices that can detect sound events.

This paper presents a complete solution for polyphonic SED task on mobile devices. The architecture includes offline training and online detection as shown in Figure 1. The offline training process involves the model training and compression. The online detection process includes acquisition of sensor data, processing of audio signals, and detecting and recording of sound events. For offline training, we will introduce the model compression method which is critical to reducing the usage of computation and storage resources. We improve the distillation method [Hinton *et al.*, 2015] to get better performance on the compact model than the complex

*Corresponding Author

models with more parameters. For online detection, we will introduce the acoustic data processing method. Finally, we will demonstrate the application how to perform sound event detection on mobile devices in near real time.

2 Architecture of the Solution

2.1 Offline Training

Getting the audio dataset, the first thing is to train the model. There are multiple models with different structures and parameters, which can detect the polyphonic sound events. However, these models are of tens of millions of parameters, like ResNet50 [He *et al.*, 2016] followed by Bi-GRUs [Cho *et al.*, 2014] or Xception [Chollet, 2017] followed by Bi-GRUs, which require high storage and computational resources to perform the detection. Model distillation method can compress the model but often cause loss of accuracy. Based on the model distillation, we propose a multi model-based distillation method for sound event detection. After training the complex models, the frame-level predictions obtained by these models are used as an extra supervision term when training the compact CRNN model. The frame-level knowledge of different models can help the compact CRNN model with fewer parameters achieve better performance. The compact CRNN model used for deployment are only of hundreds of thousands of parameters, which can reduce the storage usage and speed the forward propagation time up.

2.2 Online Detection

During the online detection process, the acoustic data is acquired from the sensor. The Pulse Code Modulation (PCM) data can be obtained by sampling and encoding acoustic data. And then the Fast Fourier Transform (FFT) is applied to the PCM data in order to get the data in the frequency domain. Besides, the data processing also includes extracting the log mel band energy features and the delta features of log mel. The delta features contain the trend information about the change of log mel band energy features, which can enrich the features and improve the model performance [Wang *et al.*, 2018]. Finally, we implement the compact CRNN as an Android library. This library can take the processed data as input and then output the detected sound events. Due to limitations of mobile devices, online detection process is performed every 10 seconds. Our results show that the library is able to detect most sound events within 10 seconds.

3 Evaluation

We evaluate the models on the DCASE 2017 Challenge Task4 dataset which is a subset of AudioSet [Gemmeke *et al.*, 2017]. The dataset consists of 17 sound events divided into two categories: "Warning" and "Vehicle". The training, testing and evaluation set contains 51172 and 488 and 1103 audio clips respectively. During the offline training process, the audio clips are re-sampled using 22.05KHz and transformed to log mel band energy features and the delta features. The segment-based instance-based average (SIA) F1 value is employed for the evaluation by using the official sed_val package [Mesaros *et al.*, 2016] with a 1s segment size. The SIA F1 values of

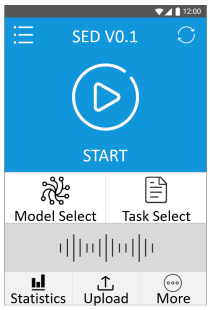

Figure 2: Snapshot of the application.

ResNet50 followed by Bi-GRUs, Xception followed by Bi-GRUs and the compact CRNN model on the evaluation set are 47.0%, 50.6% and 50.9% respectively. The compact model has better performance than complex models.

4 Demonstration

The demonstration shows the application for the polyphonic SED on mobile devices. The snapshot is shown in Figure 2. The main functions of the application include the selection of models and tasks, the detection and recording of sound events, and the uploading of labeled data. Before the detection starts, the manager must first select the corresponding model and task. At the time of detection, the application will display and record the detection result every 10 seconds. Since we use the compact model for deployment, the detection is performed in almost real time. In addition, the labeled audio data by the application can be uploaded to the server with the manager's consent, which can expand the sample database and improve the accuracy of the model by iterative training.

In our demonstration, we randomly play the audio clips in the DCASE 2017 Challenge Task4 evaluation dataset. The captured audio signal is processed and the detected sound events are presented in the screen. We compared the detection results on the application and on the server. Due to the influence of the real environment and equipment, the audio data captured by mobile devices has more noise, which makes the detection results on mobile devices often less accurate than on the server.

In conclusion, we present a SED application on mobile devices. Due to limitations of computation and storage resources on mobile devices, we use multi model-based distillation method to compress the complex models. During the detection process, the data acquired from the sensors is processed to get the log mel band energy features and delta features in order to improve the model performance. And finally, we demonstrate the application which can detect the sound events in near real time.

Acknowledgments

This work was supported by the National Grand R&D Plan (No. 2016YFB1000101) and the Science and Technology Foundation of State Key Laboratory of Sonar Technology (No. 6142109180204).

References

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014.

[Chollet, 2017] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVRP)*, pages 1800–1807. IEEE, 2017.

[Espi *et al.*, 2015] Miquel Espi, Masakiyo Fujimoto, Keisuke Kinoshita, and Tomohiro Nakatani. Exploiting spectro-temporal locality in deep learning based acoustic event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):26, 2015.

[Gemmeke *et al.*, 2017] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[Harma *et al.*, 2005] Aki Harma, Martin F McKinney, and Janto Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4 pp.–. IEEE, 2005.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVRP)*, pages 770–778. IEEE, 2016.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *CoRR, abs/1503.02531*, 2015.

[Mesaros *et al.*, 2016] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

[Parascandolo *et al.*, 2016] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE, 2016.

[Wang *et al.*, 2016] Yun Wang, Leonardo Neves, and Florian Metze. Audio-based multimedia event detection using deep recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2742–2746. IEEE, 2016.

[Wang *et al.*, 2018] Dezhi Wang, Lilun Zhang, Changchun Bao, Kele Xu, Boqing Zhu, and Qiuqiang Kong. Weakly supervised CRNN system for sound event detection with large-scale unlabeled in-domain data. *CoRR, abs/1811.00301*, 2018.

[Xu *et al.*, 2018] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In *Pacific Rim Conference on Multimedia*, pages 14–23. Springer, 2018.

[Xu *et al.*, 2019] Kele Xu, Boqing Zhu, Qiuqiang Kong, Haibo Mi, Bo Ding, Dezhi Wang, and Huaimin Wang. General audio tagging with ensembling convolutional neural network and statistical features. *The Journal of the Acoustical Society of America*, 145(3), 2019.

[Zhu *et al.*, 2018] Boqing Zhu, Kele Xu, Dezhi Wang, Lilun Zhang, Bo Li, and Yuxing Peng. Environmental Sound Classification Based on Multi-temporal Resolution Convolutional Neural Network Combining with Multi-level Features. In *Pacific Rim Conference on Multimedia*, pages 528–537. Springer, 2018.