SOUND EVENT DETECTION USING DEEP RANDOM FOREST

Chun-Yan Yu, Huang Liu, Zi-Ming Qi

College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
therica@fzu.edu.cn, liuhuang31@gmail.com, 18789468559@163.com

ABSTRACT

In this paper, we present our work on Task 3 Sound Event Detection in Real Life Audio [1]. The systems aim at dealing with the detection of overlapping audio events, where the detectors are based on deep random forest, a decision tree ensemble approach. For random forest has natural defect of detecting and classifying polyphonic events, the systems use one-vs-the-rest (OvR) multiclass/multilabel strategy, fitting one deep random forest per event class. On the development data set, the system obtained error rate value of 0.82 and F-score of 38.2%. Out of 33 submitted results, this system can rank top 14th in terms of segment error rate (0.88) and 10th in terms of F-measure (42.37%) on the evaluation data set[1].

Index Terms**—** sound event detection, polyphonic, overlapping, deep random forest, decision tree ensemble

1. INTRODUCTION

Sound event detection (SED) addresses the problem of where in time the events is happening and identification of the sound. The DCASE 2017 Task 3 Sound Event Detection in Real Life Audio task consists 6 polyphonic acoustic events, recorded in multi-source conditions similar to our everyday life. In this task, there is no control over the number of overlapping sound events at each time, which make it more challenging to handle. SED have applications in security surveillance [2], audio indexing and classification [3, 4].

In previous work, neural networks are frequently used in SED, such as Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN). Also in SED, features such as mel-frequency cepstral coefficients (MFCC), mel energy and spectrogram are widely used. MFCC and DNN as classifiers were used [5, 6]. The state of art polyphonic SED systems have used mel energy features in RNN [7, 8], and CNN trained for SED [9]. Neural networks have natural advantage to handle the overlapping events, as random forest detectors are not supposed to handle well events overlaps. For random forests, it also can be used in SED [10], Phan et al. [11] proposed discriminate decision forest detectors which are trained using both positive and negative examples.

The systems described in this report is based on deep random forest [12]. Deep random forest is a method of decision tree ensemble, which performance is very competitive to deep neural network. For detecting and classifying polyphonic events, use one-vs-the-rest (OvR) multiclass/multilabel strategy, fitting one deep random forest per event class.

The reminder of the paper is structured as follows. Section 2 describes the features used and presents a short introduction to Deep Random Forest. Section 3 presents the experimental set-up and results. Section 4 draws conclusion of our work.

2. SYSTEM DESCRIPTION

The system is based on deep random forest, built on DCASE 2016 convolutional neural network system distributed by the participants and described in [9]. Feature extraction is done with python librosa[2] package and the deep random forest modeling is based on code available on github[3]. The reminder of this section describes the core parts of system.

2.1. Features

MFCC is widely used in audio processing, such as speech recognition. Also, in DCASE 2016 task 3, most participants use MFCC. However, MFCC discard some useful information, restricts its ability for sound event detection. Mel-filter banks have been widely used for sound event detection in DCASE 2016 task 3, and have proven to be good features. In the proposed system, we use mel-filter banks.

The original 44 kHz audios are down-sampled to 16 kHz. Then, 40 filter bank features are extracted, and the audio is divided into 40ms frames with 20ms overlap using hamming window.

2.2. Deep random forest structure

A deep random forest has multi-grained scanning to scan raw features and cascade forest. In our systems, we only use the cascade forest structure. Cascade forest is inspired by the recognition of representation learning in deep neural networks, which relies mainly on the processing of raw features layer-by-layer. Figure 1 shows a cascade forest structure, each cascade level receives feature information of its previous level processing and outputs its processing results to the next level.

[1] Due to accidentally submitted wrong experimental results, but cannot anymore change official results, we withdraw our submission from the DCASE results, the results are evaluate from our corrected results.

[2] https://github.com/librosa/librosa

[3] https://github.com/pylablanche/gcForest

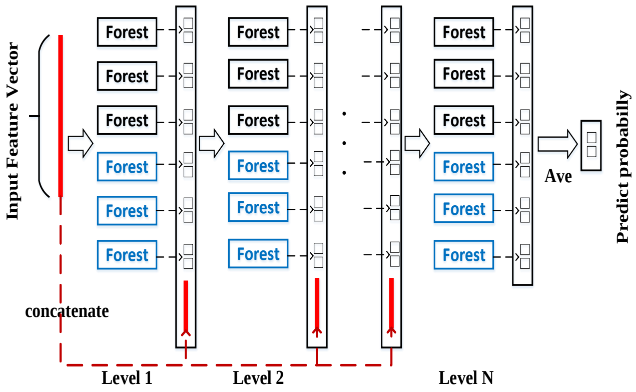

Figure 1: Cascade forest structure of deep random forest.

Each cascade of forest level is a decision tree forest ensemble. For encouraging the diversity, different types of forests, complete-random forest and random forest are used. As to each complete-random tree forest, at each node of the decision tree, a feature is selected randomly to split. Similarly, each random forest is generated by the same way, at each node of the decision tree, a subset of $\sqrt{d}$ number of features is randomly selected (d is the number of input features), and then a best *gini* value is selected to spilt.

Each forest will produces a distribution estimate of the class by calculating the percentage of the different class of training examples in the leaf nodes of the relevant instance, and then averaging all the trees in the same forest [12]. Estimate the class distribution and form a class of vector, then concatenated with the original feature vector forms a new feature vector as input to the next cascade level. In the last level of cascade, each forest will produce the class distribution. Then, average the distribution, produce the predict probability finally. The structure of deep random forest is similar to the source described in [12]. It consists of the following:

- The used features are 40 frames of log Mel filter bank acoustic features with derivations, then stack into a sequence (0.8 seconds, the length of the sequence is very long) as an input.

- Each level of cascade forest consists of 6 forest (3 complete-random forest of each contains 303 trees, 3 random forest of each contains 303 trees, the fraction of the initial samples in a node to perform a split is 0.05).

Note that the number of cascade levels can be automatically determined, because after growing a new cascade level, the performance on the validation set of the whole cascade will be estimated, if find no significant performance gain, the training procedure will terminate. Thus, the number of cascade levels is automatically determined.

2.3. One-vs-the-rest multiclass/multilabel strategy

One-vs-the-rest multiclass/multilabel strategy, also known as one-vs-all, the training strategy for each sound event class training a classifier. One advantage of this approach is its interpretability, since each class train a classifier (represented by deep random forests), which can gain the class knowledge by detecting its corresponding classifier. To apply deep random forest to polyphonic sound event detection, we use OvR multiclass/multilabel strategy, fitting one deep random forest per event class. We trained 7 deep random forest, of 6 are trained for the sound event, and 1 trained to discriminate whether the data is silence.

2.4. Event detection

The decoder was similar to the DCASE 2017 baseline system. Event probabilities are extracted from deep random forest. Then, a sliding smoothing window of 1 second length is applied. If the class is spotted within this window in more than 20 frames, the event is considered as detected in the central frame.

3. EXPERIMENTAL RESULTS

The baseline system is based on a multilayer perceptron architecture using log mel-band energies as features. Using a 5-frame context, resulting in a length of 200 of feature vector. Using these features, a neural network containing two dense layers, each layer has 50 hidden units and 20% dropout is trained for 200 epochs for each class. We also compare with a re-implementation of CNN (with two convolution layers and two fully-connected layers) [9], and a stand Random Forest (trained use OvR multiclass/multilabel strategy) with 303 trees. All experiments are conducted on 4-fold cross-validation. For the final submission, the system trained on all annotated data.

Table 1: Segment-based overall metrics for baseline multilayer perceptron system, CNN system, random forest (RF) system and the proposed model

Model	ER	F1, %
Base	0.69	56.7
CNN	0.78	51.2
RF	0.83	32.4
Ours	0.82	38.2

Table 2: Segment-based Error Rate and F-score per class for baseline multilayer perceptron system, CNN system, random forest system and the proposed model

Event	ER				F1, %			
	Base	CNN	RF	Ours	Base	CNN	RF	Ours
brakes squeaking	0.98	0.99	1.12	1.21	4.1	6.6	1.2	1.1
car	0.57	0.74	0.79	0.78	74.1	66.5	55.5	59.6
children	1.35	1.16	1.04	1.06	0.0	1.0	0.0	1.1
large vehicle	0.90	1.06	0.98	1.00	50.8	36.4	15.8	23.8
people speaking	1.25	1.33	0.99	1.07	18.5	26.4	8.9	12.7
people walking	0.84	1.07	1.01	1.11	55.6	52.5	0.0	17.9

For using error rate (ER) and F-score (F1) evaluating system performance, our model is lower than baseline and CNN systems. Similar to use GMM as classifier, deep random forest based system relies on the classifier to decide activity of sounds, the OvR multiclass/multilabel strategy is not capable of detecting onsets and offsets within the evaluated tolerance [13]. But compare to random forest based system, the model achieves 1.2% relative improvement of the ER and 17.9% relative improvement of F1.

4. CONCLUSION

A deep random forest based system submitted for DCASE 2017 challenge was described. Although better performance applications of neural networks are reported in our paper for sound event detection, we presented a new random forest ideas for SED. Use the deep random forest as classifier and trained with OvR multiclass/multilabel strategy, we achieved better performance compared to random forest. In our later experiments, found that the performance can improve by fine-tuning parameters. Future work will concentrate on using the feature sequence of shorter length after processing, and combine with multi-grained scanning of deep random forest.

5. REFERENCES

[1] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., ... & Virtanen, T. DCASE 2017 CHALLENGE SETUP: TASKS, DATASETS AND BASELINE SYSTEM.

[2] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., & Sarti, A. (2007, September). Scream and gunshot detection and localization for audio-surveillance systems. In Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on (pp. 21-26). IEEE.

[3] Cai, R., Lu, L., Hanjalic, A., Zhang, H. J., & Cai, L. H. (2006). A flexible framework for key audio effects detection and auditory context inference. IEEE Transactions on audio, speech, and language processing, 14(3), 1026-1039.

[4] Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in neural information processing systems (pp. 1096-1104).

[5] Kong, Q., Sobieraj, I., Wang, W., & Plumbley, M. D. (2016). Deep neural network baseline for DCASE challenge 2016. Proceedings of DCASE 2016.

[6] Dai Wei, J. L., Pham, P., Das, S., Qu, S., & Metze, F. (2016). Sound Event Detection for Real Life Audio DCASE Challenge. Detection and Classification of Acoustic Scenes and Events, 2016.

[7] Adavanne, S., Parascandolo, G., Pertil ä, P., Heittola, T., & Virtanen, T. (2017). Sound event detection in multichannel audio using spatial and harmonic features. arXiv preprint arXiv:1706.02293.

[8] Vu, T. H., & Wang, J. C. (2016). Acoustic scene and event recognition using recurrent neural networks. Detection and Classification of Acoustic Scenes and Events, 2016.

[9] Gorin, A., Makhazhanov, N., & Shmyrev, N. (2016). DCASE 2016 sound event detection system based on convolutional neural network. IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events.

[10] Elizalde, B., Kumar, A., Shah, A., Badlani, R., Vincent, E., Raj, B., & Lane, I. (2016). Experimentation on the DCASE challenge 2016: Task 1—Acoustic scene classification and task 3—Sound event detection in real life audio. IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events.

[11] Phan, H., Hertel, L., Maass, M., Koch, P., & Mertins, A. (2016). Car-forest: Joint classification-regression decision forests for overlapping audio event detection. arXiv preprint arXiv:1607.02306.

[12] Zhou, Z. H., & Feng, J. (2017). Deep forest: Towards an alternative to deep neural networks. arXiv preprint arXiv:1702.08835.

[13] Mesaros, A., Heittola, T., & Virtanen, T. (2016, August). TUT database for acoustic scene classification and sound event detection. In Signal Processing Conference (EUSIPCO), 2016 24th European (pp. 1128-1132). IEEE.