LETTER

Multi Model-Based Distillation for Sound Event Detection*

Yingwei FU[†,††a)], **Kele XU**[†,††], **Haibo MI**[†,††b)], **Qiuqiang KONG**[†††], **Dezhi WANG**[††††], **Huaimin WANG**[†,††], ***Nonmembers***, ***and*** **Tie HONG**[††], ***Student Member***

SUMMARY Sound event detection is intended to identify the sound events in audio recordings, which has widespread applications in real life. Recently, convolutional recurrent neural network (CRNN) models have achieved state-of-the-art performance in this task due to their capabilities in learning the representative features. However, the CRNN models are of high complexities with millions of parameters to be trained, which limits their usage for the mobile and embedded devices with limited computation resource. Model distillation is effective to distill the knowledge of a complex model to a smaller one, which can be deployed on the devices with limited computational power. In this letter, we propose a novel multi model-based distillation approach for sound event detection by making use of the knowledge from models of multiple teachers which are complementary in detecting sound events. Extensive experimental results demonstrated that our approach achieves a compression ratio about 50 times. In addition, better performance is obtained for the sound event detection task.

key words: *sound event detection, model distillation, model compression, convolutional recurrent neural network*

1. Introduction

The task of sound event detection (SED), also called as acoustic event detection aims at recognizing the onset and offset times of sound events and predicting the sound events to predefined classes. Generally speaking, SED task can be divided into monophonic sound event detection and polyphonic sound event detection. Monophonic SED indicates that the sound events are not overlapping in audio while polyphonic SED indicates that multiple sound events may occur at the same time. Currently, SED task is confronted with several challenges. Firstly, an audio clip may contain overlapping sound events, which makes it difficult to obtain representative features for detection. Secondly, an audio clip is often weakly labeled which only contains the presence of sound events without the timestamps of sound events, as it's time-consuming to label the onset and offset times of the sound events manually.

Sustainable efforts have been made to address these challenges. Traditional methods for SED rely on the shallow-architecture learners, such as Hidden Markov Models (HMMs) [1]. However, the overlap of the sound events results in difficulty to predict the sound events and timestamps. Non-negative matrix factorization (NMF) has been used to separate the overlapping sound events in an audio clip [2]. And, NMF is able to predict overlapping sounds. However, it only handles frame-level information, and ignores the temporal context. Recently, deep neural networks using log mel band energy features or mel frequency cepstral coefficients (MFCCs) features as input have shown improved performance for the SED task. In more detail, convolutional neural network (CNN) can exploit spatially local correlation across input data [3]. While, recurrent neural network (RNN) can capture long term temporal context for the audio signal [4]. By combining CNN and RNN, convolutional recurrent neural network (CRNN) has provided state-of-the-art results on various polyphonic sound event detection task.

However, most of these deep models have millions of parameters to be trained, and the models can not be applicable to the devices with limited computation and storage resources [5]. Model distillation is a general machine learning approach for model compression and performance improvement, and it uses the knowledge learned by teacher model to help the student model to train [6]. The teacher model is often a high-capacity pre-trained deep model while the student model is a small target net with lower-performance. It is worth mentioning that, for weakly-labeled SED task, the previous work [7] proposed a method of iterative training using model distillation to improve model performance.

In this letter, we propose a novel model distillation approach for polyphonic SED. The proposed approach utilizes the classification probabilities (frame-level) obtained by the teacher model as extra training supervision term for student model. Different from the iterative distillation method [7], our approach combines multiple teachers' frame-level predicted distributions to help student model improve the performance. Because different models are complementary in detecting sound events, integrating the knowledge of multiple models can let the student model have better generalization ability. We conduct experiments on the weakly labeled

Manuscript received March 26, 2019.
Manuscript revised June 24, 2019.
Manuscript publicized July 8, 2019.

[†]The authors are with National Key Laboratory of Parallel and Distributed Processing, Changsha, 410073, China.
[††]The authors are with College of Computer, National University of Defense Technology, Changsha, 410073, China.
[†††]The author is with Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK.
[††††]The author is with College of Meteorology and Oceanography, National University of Defense Technology, Changsha, 410073, China.

*This work is supported by the National Grand R&D Plan (No.2016YFB1000101) and the National Natural Science Foundation of China (No.61806214).

a) E-mail: yingwei_fu_nudt@163.com
b) E-mail: haibo_mihb@126.com (Corresponding author)

DOI: 10.1587/transinf.2019EDL8062

Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge Task 4 dataset. We find that our approach can provide a model compression ratio about 50 times, and the student model can get better performance.

This letter is organized as follows. Section 2 descries the CRNN framework. Section 3 presents the proposed method. Section 4 shows the experiments. Conclusion is given in Sect. 5.

2. Framework

In our experiment, we employ the CRNN for the task, which provides state-of-the-art performance for SED [8]. The framework of CRNN is shown in Fig. 1. The extracted log mel band energy and delta features of log mel are combined as the input for the CNN. As demonstrated in previous study [5], the delta features can improve the performance as it may enrich the input features dimension. Then the output of the convolutional layers are employed as input to two bidirectional RNNs. The RNN component is followed by a gate structure and an attention layer [8] so that the information related to the temporal sequence can be retained to improve the performance.

As a proof of concept, the CNN components selected can be ResNet50 [9], Xception [10], DenseNet201 [11] and Inception-V3 [12], as these models have shown to be effective in classification tasks. Two bidirectional GRUs [13] with different activation functions are adopted to capture the temporal information, and each network consists of 128 cells. The GRUs outputs are then element-wise multiplied by the gated structure. The attention layer includes two independent fully connected (FC) layers whose activation functions are softmax and sigmoid respectively. The FC layer with sigmoid activation function predicts the probability of sound events at each frame. The FC layer with softmax activation function is used to attend to the frames that may contain sound events. The output of the FC layer with sigmoid activation function is the frame-level prediction which is denoted as $Y_f \in [0,1]^{T\times K}$. The probability of sound event j at frame i is obtained by:

$$Y_f(i,j) = \sigma(X_{i,j}) \tag{1}$$

where $\sigma(X_{i,j}) = \frac{1}{1+e^{-X_{i,j}}}$, and $X \in \mathbb{R}^{T\times K}$ is the output matrix of FC layer before activation function. T is the number of frames and K is the number of sound event types. The clip-level prediction corresponding to each audio is defined as $Y_c \in [0,1]^K$ and the probability of sound event j is obtained by:

$$Y_c(j) = \frac{\sum_{i=1}^{T} \phi(X_{i,j}) \odot \sigma(X_{i,j})}{\sum_{i=1}^{T} \phi(X_{i,j})} \tag{2}$$

where $\phi(X_{i,j}) = \frac{e^{X_{i,j}}}{\sum_{k=1}^{K} e^{X_{i,k}}}$ is the output of the FC layer with softmax activation function, and $\odot$ is the element-wise multiplication. Since the training data of the DCASE 2017 Challenge Task 4 dataset is weakly labeled, we use Eq. (2) to get the clip-level prediction and the binary cross-entropy between the clip-level prediction Y_c and the ground truth label of the audio is used as training loss. The training loss is defined as:

$$L_b = -\sum_{i=1}^{N} ({Y_c}^i \log Y^i + (1-{Y_c}^i)\log(1-Y^i)) \tag{3}$$

where ${Y_c}^i$ and Y^i denote the estimated clip-level prediction vector and ground truth label vector at sample index i, respectively. The batch size is N.

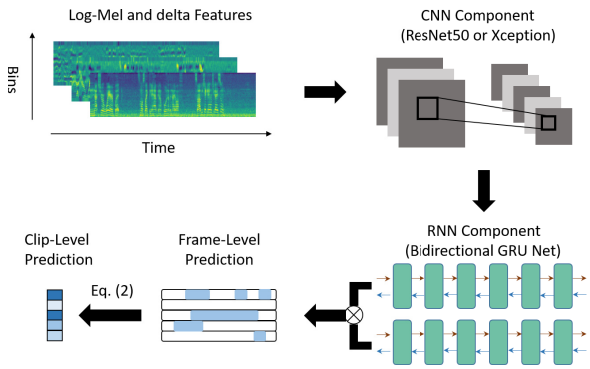

Fig. 1 Framework of CRNN for weakly labeled polyphonic SED task.

3. Model Distillation

Model distillation involves transferring knowledge from a complex model (teacher model) to a smaller one (student model) [6]. The basic principle of distillation is to introduce additional supervision of the teacher model in student model training, beyond the conventional supervised learning objectives. The implementation of our approach is shown in Fig. 2. As SED task is required to determine the onset and offset times of sound events, frame-wise distillation can transmit the knowledge of temporal information in frames instead of the whole clip. Indeed, as the choice of the teacher model can be diverse, distillation from multiple teacher models can enhance the generalization ability of student model and achieve better performance.

3.1 Teacher Model and Student Model

The CRNN models proposed in Sect. 2 provide state-of-the-art performance for SED. However, there are millions of parameters in the models which are not applicable for the mobile and embedded devices with limited resources. These models are typical teacher models.

The amount of CRNN framework parameters mainly depends on the complexity of CNN and RNN components. In our experiment, for the student model, the CNN component is a compact CNN which consists of three blocks, and each block includes one convolution layer followed by one batch normalization layer, one gate layer and one max-pooling layer. The gate layer [8] in compact CNN consists

of a sigmoid branch and a linear branch. The RNN component in student model is two bidirectional GRUs which consists of 80 cells respectively. The parameter of the gate structure and the attention layer is only a small part of the entire CRNN framework and both of them are helpful in improving the performance, so they are adopted in student model.

3.2 Frame-Wise Distillation

The Kullback-Leibler Divergence (KLD) is often used for training the student model to imitate the prediction of the teacher model. In frame-wise distillation for SED task, we apply distillation to each frame. And the frame-wise KLD loss is defined as below:

$$L_f = -\frac{\tau^2 \sum_{i=1}^{T} \sum_{j=1}^{K} \lambda_t(i,j) \log \lambda_s(i,j)}{T} \tag{4}$$

where τ is a temperature hyper-parameter which controls the soft degree of probability distribution. The symbols λ_t and λ_s are frame-wise soft labels from the teacher model and student model. Compared with the approach which uses the frame-level prediction directly, the soft label can be more representative. The λ_t and λ_s are defined as below:

$$\begin{aligned} \lambda_t(i,j) &= \frac{e^{\upsilon(Y_{tf}(i,j))}}{\sum_{k=1}^{K} e^{\upsilon(Y_{tf}(i,k))}} \\ \lambda_s(i,j) &= \frac{e^{\upsilon(Y_{sf}(i,j))}}{\sum_{k=1}^{K} e^{\upsilon(Y_{sf}(i,k))}} \end{aligned} \tag{5}$$

where $\upsilon(x) = \log \frac{x}{\tau - \tau x}$, and it converts the classification result into logit form with temperature. $Y_{tf}(i,j)$ and $Y_{sf}(i,j)$ are the frame-level probability of sound event j at frame i in teacher model and student model respectively.

3.3 Multi Model-Based Distillation

Although frame-wise distillation can help student model improve the performance, it only utilizes the knowledge of one teacher model. However, we argue that the students often have multiple teachers to obtain knowledge. In this part, we explore multi model-based distillation which makes use of the information from multiple teacher models.

We denote the frame-wise KLD loss of different teacher models as L_f^i, $i \in \{1, \ldots, M\}$, and the number of teacher models is M. The mixed KLD loss is defined as below:

$$L_m = \sum_{i=1}^{M} \theta_i L_f^i \tag{6}$$

where θ_i is the weight of each teacher model. We set $\theta_i = \frac{1}{M}$, $i \in \{1, \ldots, M\}$ which denotes average mixing. The mixed teacher models can encourage the student model to behave linearly between teacher models and enhance the generalization ability.

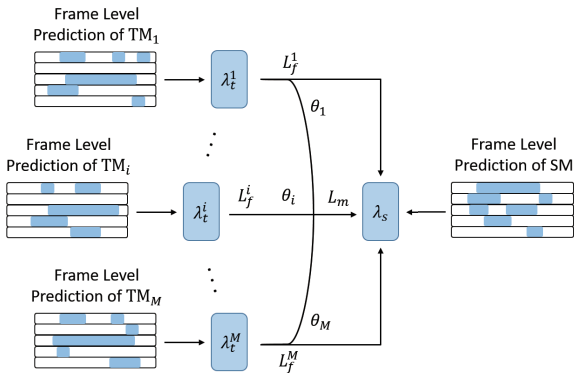

Fig. 2 Frame-wise distillation and multi model-based distillation schematic. The frame-level prediction is transformed into soft label and the KLD between these labels is used as extra loss.

4. Experiments

4.1 Experimental Settings

Experiments are conducted on the DCASE 2017 Challenge Task 4 dataset. The training, testing and evaluation set contains 51172 and 488 and 1103 audio clips respectively. The training set is weakly labeled. For testing and evaluation, strong labels with timestamps are provided. We re-sample audio clips using 22.05KHz and transform the wave-form to log mel band energy. Then the first and second order delta features of log mel features are used as the other two channels features to form a 3-channel features as input, and each channel feature has the size of 320×128. All the methods use the same input features.

The structure of the baseline CNN is the same as that of DCASE 2017 Challenge baseline system. The CNN components (ResNet50, Xception, DenseNet201 and Inception-V3) of the teacher CRNN models are fine-tuned from the pre-trained models which are trained on the ImageNet dataset. And the RNN components of the teacher models and all components of the student models are fully trained using DCASE Challenge dataset. The F1 value measure is employed for the evaluation by using the official sed_val package [14] with a 1s segment size and a 200ms collar on onsets and a 200ms/20% of the events length collar on offsets.

4.2 The Influence of the Number of Teacher Models

In our experiments, we firstly study the influence of the number of teacher models. The teacher models include ResNet50, Xception, DenseNet201 and Inception-V3 which are followed by two bidirectional GRUs consisting of 128 cells respectively, a gate structure and an attention layer. The amount of model parameters are 25.36 million, 21.46 million, 18.90 million and 22.21 million respectively. For the combinations of different numbers of teacher models, we test all types of combinations and report the average F1 values. Figure 3 gives the quantitative comparison. We find increasing the number of the teacher models does not pro-

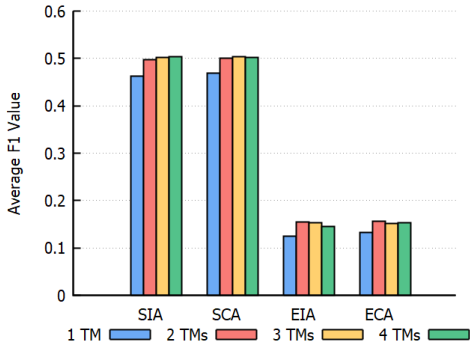

Fig. 3 The average F1 value of the student models (with different settings of the teacher models). The SIA, SCA, EIA and ECA denotes the segment-based instance-based average F1 value, the segment-based class-based average F1 value, the event-based instance-based average F1 value and the event-based class-based average F1 value respectively.

Table 1 The components and the amount of parameters for CRNN frameworks.

	CNN component	RNN component	CRNN parameters
TM1	ResNet50	128 cells Bi-GRU	25.36 million
TM2	Xception	128 cells Bi-GRU	21.46 million
SM	Compact CNN	80 cells Bi-GRU	0.48 million

vide dramatic performance gain but increasing the computation cost, and 2 teacher models can provide satisfied performance. In the following letter, we use 2 teacher models for the model distillation.

4.3 Model Compress Ratio

The configuration of CRNN frameworks is shown in Table 1. As can be seen from the table, compared with teacher models (TM1 and TM2), our method can provide a compression ratio of 53× and 45× for the student model (SM). The student model with reduced model size can be deployed on the mobile or embedded devices with limited resources.

4.4 Model Performance

The results of F1 value comparisons on the evaluation set are shown in Table 2. Owing to using our 3-channel features as input, the SIA of baseline CNN increased slightly (compared with 28.4% in DCASE baseline system). Compared with the baseline, the CRNN models have better performance. Frame-wise distillation and multi model-based distillation methods can both improve the F1 value of SM, and the multi model-based distillation shows superior improvement versus frame-wise distillation because it mixes the knowledge of multiple teacher models. The SIA, EIA and ECA of the SM with L_m are 50.9%, 16.8% and 17.1% which are the best result among all the methods. Although the SCA of the SM with L_m is 2.8% lower than the F1 value in TM2, it improves remarkably compared with the F1 value in SM.

Table 2 The results of F1 value on the evaluation set.

Method	SIA	SCA	EIA	ECA
Baseline CNN	29.5%	30.7%	4.1%	5.0%
TM1	47.0%	48.9%	11.9%	12.5%
TM2	50.6%	**53.3**%	14.4%	15.5%
SM	40.4%	43.0%	10.7%	12.3%
SM with L_f^1	44.6%	46.2%	13.5%	14.3%
SM with L_f^2	46.6%	49.6%	14.5%	15.2%
SM with L_m	**50.9**%	50.5%	**16.8**%	**17.1**%

5. Conclusion

This letter proposes a novel multi model-based distillation method on CRNN for weakly labeled polyphonic SED task. The proposed method can utilize the advantages of different models to improve the performance of the student model, and the student model with less parameters is applicable to the devices with limited computation and storage resources.

References

[1] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," IEEE 2010 18th European Signal Processing Conference, pp.1267–1271, 2010.
[2] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," ICASSP, pp.8677–8681, IEEE, 2013.
[3] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," EURASIP Journal on Audio, Speech, and Music Processing, vol.2015, no.1, p.26, 2015.
[4] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," ICASSP, pp.6440–6444, IEEE, 2016.
[5] D. Wang, L. Zhang, C. Bao, K. Xu, B. Zhu, and Q. Kong, "Weakly supervised CRNN system for sound event detection with large-scale unlabeled in-domain data," CoRR, abs/1811.00301, 2018.
[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," CoRR, abs/1503.02531, 2015.
[7] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Iterative knowledge distillation in R-CNNs for weakly-labeled semi-supervised sound event detection," tech. rep., DCASE2018 Challenge, 2018.
[8] Y. Xu, Q. Kong, W. Wang, and M.D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," ICASSP, pp.121–125, IEEE, 2018.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CVPR, pp.770–778, IEEE, 2016.
[10] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," CVPR, pp.1800–1807, IEEE, 2017.
[11] G. Huang, Z. Liu, L.V.D. Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," CVPR, pp.4700–4708, 2017.
[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," CVPR, pp.2818–2826, 2016.
[13] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," EMNLP, pp.1724–1734, Association for Computational Linguistics, 2014.
[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," Applied Sciences, vol.6, no.6, p.162, 2016.