MULTI-SCALE CONVOLUTIONAL RECURRENT NEURAL NETWORK WITH ENSEMBLE METHOD FOR WEAKLY LABELED SOUND EVENT DETECTION

Yingmei Guo[1], *Zhihao Ouyang*[1], *Mingxing Xu*[1,*], *Zhiyong Wu*[1], *Jianming Wu*[2], *Bin Su*[1]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

[2]KDDI Research, Inc. Fujimino-shi, Japan
{guoym18, oyzh18,sub18}@mails.tsinghua.edu.cn, xumx@tsinghua.edu.cn,
{zywu}@se.cuhk.edu.hk , ji-wu@kddi-research.jp

ABSTRACT

In this paper, we describe our contributions to the challenge of detection and classification of acoustic scenes and events. We propose multi-scale convolutional recurrent neural network(Multi-scale CRNN), a novel weakly-supervised learning framework for sound event detection. By integrating information from different time resolutions, the multi-scale method can capture both the fine-grained and coarse-grained features of sound events and model the temporal dependency including fine-grained dependency and long-term dependency. Furthermore, the ensemble method proposed in the paper reduces the frame-level prediction errors with classification results. The proposed method achieves 29.2% in the event-based F1-score and 1.40 in event-based error rate in development set of DCASE2018 task4 compared to the baseline of 14.1% F-value and 1.54 error rate [1].

Index Terms— sound event detection, deep learning, convolutional recurrent neural network, multi-scale model, ensemble model

1. INTRODUCTION

The perception and understanding of sound play an important role in human interaction with the surroundings. With the continuous development of smart homes, auto-driving cars and security surveillance devices, sound event detection has received increasing attention. With the development of multimedia and network technologies, the scale of audio data grows rapidly. Therefore, how to effectively identify, label and retrieve useful content from audios has become an urgent problem to be solved.

The target of the task is to provide not only the event class but also the event time boundaries given that multiple events can be present in an audio recording [2]. Many methods can be applied to the sound event detection task, such as Gaussian Mixture Model(GMM)[3], Hidden Markov Model(HMM)[4], non-negative matrix factorization (NMF)[5] and Deep Neural Network(DNN)[6]. [7] proposed a system using the entire audio clip and the segmented audio clip separately as the input to predict audio events in a short-time segment. Moreover, convolutional neural network structures such as AlexNet[8], VGG[6], Inception[9] and ResNet[10] also perform well in sound event detection[11]. [12] proposed multi-scale RNN to balance the modeling of both the fine-grained and long-term dependency.

As for the weakly labeled sound event detection, the weakly labeled data lacks frame-level strong labels. In [13], by training the classification of the entire audio, the frame-level prediction could be used as detection result and achieve weakly supervised learning without strong frame labels. [14] applied multi-instance learning(MIL) where each audio was regarded as a packet, frames or short segments are regarded as examples. [15] used the model proposed in [16] to do classification and applied transposed convolutional network to reconstruct the signal of original audio then make frame-level prediction. However, a significant amount of research is still needed to reliably detect sound events in realistic soundscapes, where multiple sounds are present simultaneously and the target sounds usually mix with environment noise.

In this paper, we propose multi-scale convolutional recurrent neural network. First, the CNN structure is proposed by [13] which applies the learnable gated linear units(GLUs)[11] to replace the ReLU[17] activation. This learnable gate is able to select the most related features corresponding to the audio labels. The RNN structure followed the CNN can model the temporal dependency. Second, the multi-scale method is applied to capture useful information from both the fine-grained and coarse-grained features of sound events and balance the modeling of both the fine-grained and long-term dependency. Finally, the ensemble method can further help to reduce the frame-level prediction errors with classification results since identifying the sound events occurred in the audio is much easier than predicting the event time boundaries. Section 2 describes our multi-scale CRNN architecture and the ensem-

* Corresponding author

ble method. Section 3 shows and discusses the experiments and results. In the end, section 4 summarizes this paper and mentions our future work.

2. PROPOSED METHOD

2.1. Network Architecture

The main model structure is shown in Fig.1. The multi-scale method used in the model combines two CRNNs separately work at fine-scale and coarse-scale. The fine-grained input and the coarse-grained input of network are features described in the section 3.2 with the shape of $(N, 1200, 64)$ and $(N, 240, 64)$ separately. N is the number of audios, 64 is the number of mel-bins and the second dimension of feature arrays is the number of frames.

First, the fine-grained feature sequence X_{fine} is split into five subsequences X_{fine_1}, X_{fine_2}, X_{fine_3}, X_{fine_4}, X_{fine_5} and feed into the fine-scale CRNN with shared parameters. At the same time, the coarse-grained feature sequence X_{coarse} feed into coarse-scale CRNN. Then we combine the extracted features from fine-scale CRNN and coarse-scale CRNN by concatenation: at each fine-scale time step, we replicate the corresponding vector of coarse-scale CRNN's output and concatenate it with feature vector of the fine-scale time step. The multi-scale CRNN which utilize the fine-grained and coarse-grained features of sound events together can model both fine-grained and long-term dependency simultaneously. The concatenation process is defined as:

$$O^t = concatenate(O^t_{fine}, O^{\lfloor \frac{t}{5} \rfloor}_{coarse}) \tag{1}$$

where O^t_{fine},$O^{\lfloor \frac{t}{5} \rfloor}_{coarse}$ are the output of fine-scale CRNN at time t and coarse-scale CRNN at time $\lfloor \frac{t}{5} \rfloor$ separately. It should be noted that we only do convolution and pooling operations on spectral axis to keep the time resolution of the input. O^t is the feature vector of audio clip at fine-scale time step t integrating information from different time resolutions. Each cell of the coarse-scale CRNN interacts with five cells of the fine-scale CRNN by concatenation.

Another question is about how to combine coarse-grained and fine-grained features. Intuitively, for a specific range of audio features, we can replicate course-grained feature k times, then use self attention mechanism to concatenate them with k fine-grained features:

$$A = softmax(W_1 tanh(W_2 O^T_{fine})) \tag{2}$$

$$O_{fine'} = AO_{fine} \tag{3}$$

where O_{fine} is the output of fine-scale CRNN.

From our experiment, self attention mechanism brings trivial improvement versus simple all-one attention matrix (equivalent to simple concatenate). Thus we replicate the coarse-grained feature k times and simply concatenate them with k fine-grained features.

After concatenation, final probabilistic predictions are made at fine-scale time step using a fully connected layer with sigmoid output:

$$Y^t = \sigma(UO^t + c) \tag{4}$$

where Y^t is probabilistic prediction at time step t of the audio. As no frame-level strong labels are provided, the temporal information extracted from each occurring sound event in the audio can only be inferred as intermediate variables. Then we average probabilistic predictions Y^t of all frames to determine the existence of the event in the audio clip.

The training is supervised by minimizing the binary cross-entropy loss:

$$loss(Y, \hat{Y}) = \hat{Y} log Y + (1 - \hat{Y}) log(1 - Y) \tag{5}$$

where Y,$\hat{Y}$ denote the predicted probability and ground truth of an audio recording, respectively.

The CNN structure in the model is proposed by [13] which applies GLUs to introduce the attention mechanism to all layers of the neural network. GLUs are defined as:

$$y = (W * x + a) \odot \sigma(V * x + b) \tag{6}$$

where x is the input of the first layer or the feature maps of the interval layers of CNN. GLUs can attend to the T-F bin with related audio events by setting its value close to one otherwise close to zero and control the information passed in the hierarchy. The RNN structure in the model is bidirectional to learn useful contextual information from both time directions. Fig.2 shows the CRNN structure.

2.2. Ensemble Method

As the audio clips are weakly labeled, it is easy to do classification instead of detection. Therefore, it is obvious the classification task can achieve higher accuracy. In our ensemble method, classification results can help to reduce the frame-level prediction errors. When the event occurs in a frame it must occur in that audio clip. Sometimes proposed multi-scale CRNN gives the result that a sound event occurs in a frame while the ensemble model shows that sound event doesn't occur in the audio clip. We correct the multi-scale CRNN's prediction of that frame because of the higher reliability of ensemble model.

We use the CNN based model introduced in [7], single-scale CRNN[13] and multi-scale CRNN proposed by our paper to predict the existence of sound events. Then we fuse the results of aforementioned three models to produce the final prediction. Fig.3 shows the structure of the CNN based model.

We get final classification results using majority vote method. Every model makes a prediction for each audio and the final output prediction is the one that receives more than

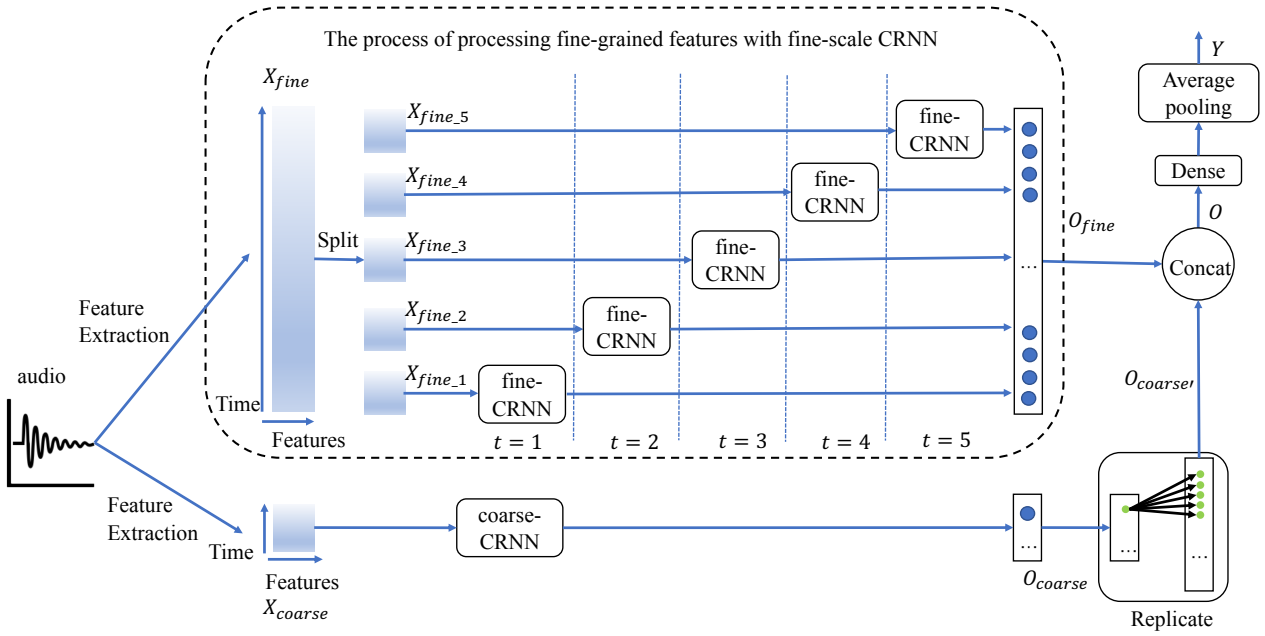

Fig. 1. Multi-scale CRNN: During computation, the fine-grained input sequence is splits into five subsequences X_{fine_1}, X_{fine_2}, X_{fine_3}, X_{fine_4}, X_{fine_5} and then feed into the CRNN with shared parameters and the coarse-grained input sequence feed into another CRNN. Each cell of the coarse-scale CRNN interacts with five cells of the fine-scale CRNN by concatenation.

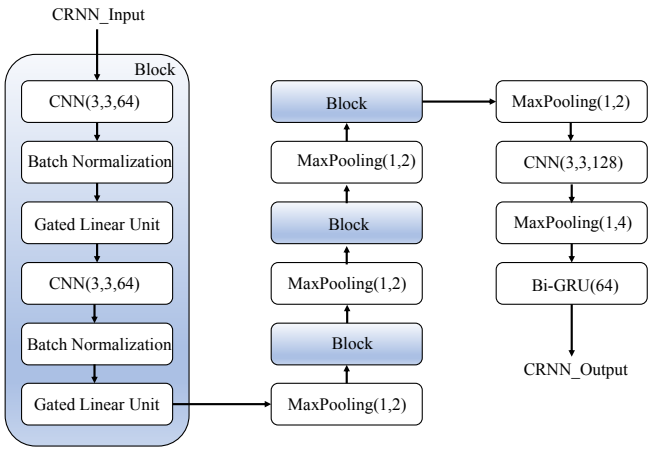

Fig. 2. CRNN structure. The left part describes the details of the "Block".

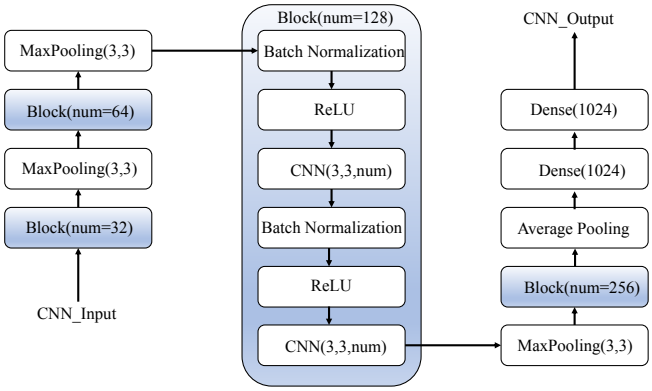

Fig. 3. CNN based classification model architecture. The center part describes the details of the "Block" and the number in the brackets of the "Block" is the number of filters of convolutional layers.

half of the votes. The method is defined as:

$$R_{ij} = \begin{cases} 1, r1_{ij} + r2_{ij} + r3_{ij} \geq 2 \\ 0, r1_{ij} + r2_{ij} + r3_{ij} < 2 \end{cases} \quad (7)$$

where R_i is the final output prediction of the i-th audio, $r1_i$, $r2_i$, $r3_i$ are the predictions of the i-th audio with the single scale model, multi-scale model and CNN model separately. $R_{ij} = 1$ when more than two models give the same prediction: j-th sound event occurs in the i-th audio. Otherwise $R_{ij} = 0$. We use the classification models as sound event detectors and correct the frame-level prediction errors.

3. EXPERIMENTS

3.1. Task and Data

The target of task is to provide not only the event class but also the event time boundaries given that multiple events can be present in an audio recording.

The data are YouTube videos excerpt from domestic context. We focus on a subset of Audioset which consists of 10 classes of sound events: Speech, Dog, Cat, Alarm, Dishes, Frying, Blender, Running Water, Vacuum cleaner and Electric shaver [18].

The labeled training set contains 1578 clips(2244 class occurrences) for which weak annotations have been verified and cross-checked. The test set contains 288 clips(906 events) annotated with strong labels, with timestamps(obtained by human annotators).

3.2. Set-up

Log-Mel filter banks are used as our features. In general, we resample all audios to 16kHz and calculate the mel-spectrograms with 64 mel-bins at two scales with hop sizes of 0.0415 seconds(coarse-scale) and 0.0083 seconds(fine-scale). The window size of the two scales for short-time Fourier transform is 0.064 seconds. Then the mel-spectrogram is converted into logarithmic scale and standardized by subtracting the mean value and dividing by the standard deviation. There are some audios that are shorter than 10 seconds, and the features extracted from the audios are zero-padded to equalize the length.

In the training phase, we apply the binary cross-entropy loss between the predicted probability and ground truth of an audio recording. Adam[19] is used as the stochastic optimization method.

3.3. Results

The results of audio tagging and weakly supervised sound event detection will be given in this section.

3.3.1. Audio Tagging

Table 1 shows the Precision, Recall and F1-value of multiple different systems on development set of DCASE2018 task4. We can find that multi-scale CRNN model(F1=85.5%) is better than single scale CRNN(F1=82.0%). We also fuse the three models by majority voting. The fusion model achieves the best score(F1=88.7%).

3.3.2. Sound event detection

Table 2 shows the F1-value and error rate of multi-scale CRNN using and not using classification results for correction. It shows that post-processing of the frame-level predictions is important and can improve the system performance by 6.1%. Submissions are evaluated with event-based measures with a 200ms collar on onsets and a 200ms collar on offsets.

Table 3 shows the F1-value and the error rate of single-scale CRNN, multi-scale CRNN and the baseline on development set of DCASE2018 task4. Multi-scale CRNN has the best performance. It demonstrates that multi-scale method can capture useful information and produce better results.

Table 1. Comparison of multi-scale CRNN, single-scale CRNN, CNN based model and fusion model on audio tagging.

Models	Precision(%)	Recall(%)	F1-value(%)
CNN based model[7]	85.1	85.1	85.1
Single-scale CRNN[13]	83.2	80.9	82.0
Multi-scale CRNN	83.5	87.6	85.5
Fusion	87.7	89.8	88.7

Table 2. Comparison of F1-value and the error rate of multi-scale CRNN using and not using classification results for correction on sound event detection.

Models	F1-value(%)	Error rate(%)
Multi-scale CRNN not using correction	23.1	1.90
Multi-scale CRNN using correction	29.2	1.40

4. CONCLUSIONS

In this paper, we propose multi-scale convolutional recurrent neural network. The CNN structure in the model applies the learnable gated linear units to control the information flow to the next layer. The RNN structure followed the CNN can model the temporal dependency. The multi-scale method is applied to capture useful information from both the fine-grained and coarse-grained features of sound events. It also balances the modeling of both the fine-grained and long-term dependency. The ensemble method can help to reduce the frame-level prediction errors with classification results.

We also tried several methods to improve the system using unlabeled data, and we plan to further improve the system through this idea. Our Future work can be done by exploring the possibility to exploit a large amount of unlabeled and unbalanced training data together with a small weakly annotated training set.

Table 3. Comparison of multi-scale CRNN, single-scale CRNN, and baseline on sound event detection.

Models	F1-value(%)	Error rate(%)
Baseline [1]	14.1	1.54
single-scale CRNN[13]	24.4	1.24
Multi-scale CRNN	29.2	1.40

5. REFERENCES

[1] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[2] Anurag Kumar and Bhiksha Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 1038–1047.

[3] Gregoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, and Axel Roebel, "A morphological model for simulating acoustic scenes and its application to sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1854–1864, 2016.

[4] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1, 2013.

[5] Onur Dikmen and Annamaria Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[6] Ting-Wei Su, Jen-Yu Liu, and Yi-Hsuan Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 791–795.

[7] Donmoon Lee, Subin Lee, Yoonchang Han, and Kyogu Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[10] 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*, 2016, pp. 770–778.

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[12] Rui Lu, Zhiyao Duan, and Changshui Zhang, "Multi-scale recurrent neural network for sound event detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 131–135.

[13] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

[14] Justin Salamon, Brian McFee, Peter Li, and Juan Pablo Bello, "Dcase 2017 submission: Multiple instance learning for sound event detection," Tech. Rep., Technical report, DCASE2017 Challenge (September 2017), 2017.

[15] Szu-Yu Chou, SR Jang, and Yi-Hsuan Yang, "Framecnn: a weakly-supervised learning framework for frame-wise acoustic event detection and classification," *Recall*, vol. 14, pp. 55–4, 2017.

[16] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[18] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.