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

Yingmei Guo, Mingxing Xu

Tsinghua University
Department of Computer Science and Technology, 30 ShuangQing Road
HaiDian, Beijing 100084, China
qniguoym@163.com

Jianming Wu, Yanan Wang, Keiichiro Hoashi

KDDI Research, Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502
Japan
{ji-wu, wa-yanan, hoashi}@kddi-research.jp

ABSTRACT

In this paper, we describe our contributions to the challenge of detection and classification of acoustic scenes and events 2018(DCASE2018). 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 dependencies including fine-grained dependencies and long-term dependencies. CRNN using learnable gated linear units(GLUs) can also help to select the most related features corresponding to the audio labels. Furthermore, the ensemble method proposed in the paper can help to correct the frame-level prediction errors with classification results, as identifying the sound events occurred in the audio is much easier than providing the event time boundaries. 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, Weakly-supervised learning, Deep learning, Convolutional recurrent neural network, Multi-scale model

1. INTRODUCTION

Recently a large-scale weakly supervised sound event detection task of DCASE2018 challenge was proposed[2]. 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[3]. The task employs a subset of "Audioset: An Ontology And Human-Labeled Dataset For Audio Events" by Google[4].

Sound is one of the signals carrying information. The perception and understanding of sound plays 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, audio data grows rapidly in the databases. Therefore, how to identify, label and retrieve useful content from audios effectively has become an urgent problem to be solved.

Many methods can be applied in the sound event detection task, such as Gaussian Mixture Models(GMM)[5], Hidden Markov Models(HMM)[6], non-negative matrix factorization(NMF)[7] and Deep Neural Network(DNN)[8]. ConvNets showed the promising results in a large number of computer vision tasks and have been actively adopted for audio content analysis[8]. [9] proposed a sound event detection system that combines global-input model and separated-input model using the entire and a segmented audio clip as input separately to predict audio events in a short-time segment. Moreover, convolutional neural network structures that perform well in image recognition tasks such as AlexNet[10], VGG[11], Inception[12] and ResNet[13] also perform well in sound event detection[14]. Eghbal-Zadeh et al. used the VGG classifier in the DCASE2016 acoustic scene recognition task and ranked first [15]. [16] proposed multi-scale RNN to balance the modeling of both the fine-grained and long-term dependencies.

As for the weakly labeled sound event detection, the weakly labeled data lacks frame-level strong labels. In [17], the frame-level prediction information was used as an intermediate variable, which can influence the final output of the model and be weakly supervised. [18] applied multi-instance learning(MIL) in sound event detection where each audio is regarded as a packet and frames or short segments are regarded as examples. [19] used the model proposed in [20] to do classification and applied transposed convolutional network to reconstruct the signal of original audio and make frame-level predictions. However, a significant amount of research is still needed to reliably detect sound events in realistic soundscapes, where multiple sounds are present, often simultaneously, and distorted by the environment.

In this paper, we propose multi-scale convolutional recurrent neural network. The CNN structure in the model is proposed by [17] which applies the learnable gated linear units(GLUs)[21] to replace the ReLU[22] activation after each layer of convolutional neural network. This learnable gate is able to control the information flow to the next layer. The RNN structure followed the CNN can model the temporal dependencies. 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. The ensemble method can further help to correct the frame-level prediction errors with classification results. Section 2 describes the our multi-scale CRNN architecture. Section 3 shows and discusses the experiments and results. In the end, section 4 summarizes and plans for the future work.

2. PROPOSED METHOD

2.1. Network Architecture

The main model structure is shown in Fig.1. The fine-grained input and the coarse-grained input of network are features described in the section 3.2 with the shape of (1,1200,64) and (1,240,64) separately. Some chunks extracted from audios with length shorter than 10 seconds are zero-padded to equalize the length.

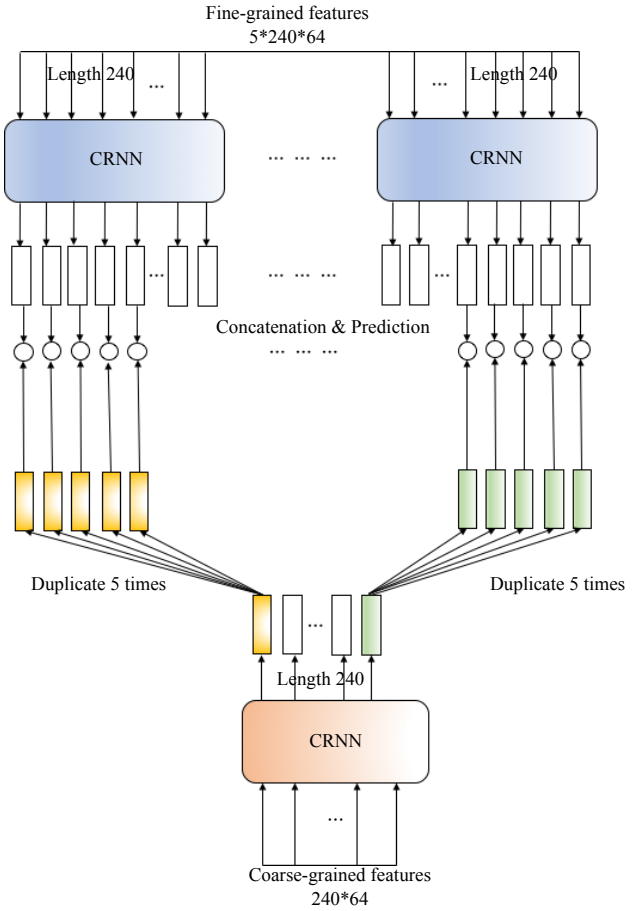

Figure 1: Multi-scale CRNN: Fine-grained feature sequence has a length of 1200 and coarse-grained feature sequence has a length of 240. During computation, the fine-grained sequence is splits into five subsequences to feed into the same CRNN structure with same parameters.

The CNN structure in the model is proposed by [17] which applies the learnable gated linear units(GLUs)[21] to replace the ReLU[22] activation after each layer of convolutional neural network. The motivation of using GLUs in audio classification is to introduce the attention mechanism to all layers of the neural network. GLUs are defined as:

$$Y = (W * X + a) \odot \sigma(V * X + b) \quad (1)$$

where X is the input of the first layer or the feature maps of the interval layers, W and V are the convolutional filters, a and b are the biases, σ is the sigmoid non-linearity, ⊙ is the element-wise product and * is the convolutional operator. It 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 on 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.

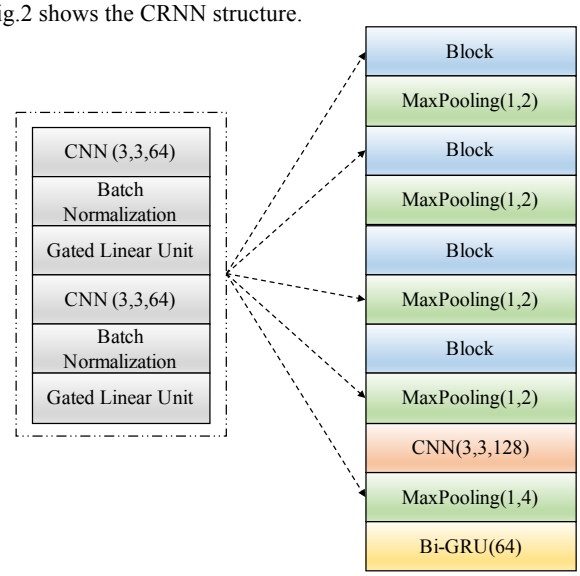

Figure 2: CRNN structure. The left part describes the details of the "Block".

The multi-scale method used in the model combines two CRNNs separately work at fine-scale and coarse-scale. The multi-scale CRNN can 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. The last layers of multi-scale models are aligned so that each cell of the coarse-scale CRNN interacts with five cells of the fine-scale CRNN by concatenation.

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

$$y_t = \sigma\left(W\left(h1_t, h2_{\lfloor \frac{t}{5} \rfloor}\right) + b\right) \quad (2)$$

Where $h1_t, h2_{\lfloor \frac{t}{5} \rfloor}$ are the output of fine-scale CRNN at time t and coarse-scale CRNN at time $\left\lfloor \frac{t}{5} \right\rfloor$ separately. It should be noted that we only do convolution and pooling operations on spectral axis to keep the time resolution of the input.

As no frame-level strong labels are provided, the temporal information of each occurring sound event in the audio can only be weakly supervised inferred as intermediate variables. We aggregate probabilistic predictions of all frames to determine the existence of the event in that audio clip.

2.2. Ensemble Method

As the audio clips are weakly labeled, it is easy to do classification instead of detection. As a result of that, it is obvious the classification task can achieve higher accuracy. The ensemble method can help to correct the frame-level prediction errors with classification results.

We use the CNN based model introduced in [9] and a single-scale model similar with the model proposed in the paper to do classification and fuse the results of three models to produce the final predictions. Fig.3 shows the structure of the CNN based model.

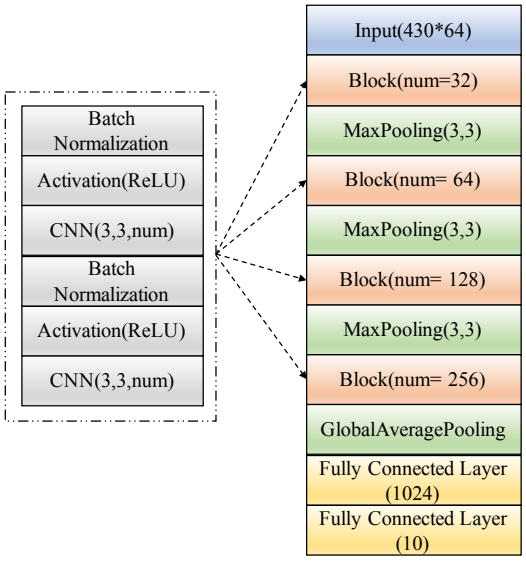

Figure 3: CNN based classification model architecture. The left part describes the details of the "Block" and the number in the brackets of the "Block" is the number of filters of convolutional layers.

We get final classification results using fusion method. The method is defined as:

$$R_i = \begin{cases} 1, r1_i + r2_i + r3_i \geq 2 \\ 0, r1_i + r2_i + r3_i < 2 \end{cases} \quad (3)$$

where R_i is the label of the i-th audio, $r1_i$, $r2_i$, $r3_i$ are the labels of the i-th audio with the single scale model, multi-scale model and CNN model separtely.

We use the classification models as sound event detectors and correct the frame-level prediction errors. When the event occurs in a frame it must occur in that audio clip.

3. EXPERIMENTS

3.1. Task and Data

The target of the DCASE2018 task4 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 are focused on a subset of Audioset that consists of 10 classes of sound events: Speech, Dog, Cat, Alarm, Dishes, Frying, Blender, Running Water, Vacuum cleaner and Electric shaver.

3.2. Set-up

Log-Mel filter banks are used as our features. In general, we resample all audios to 44.1kHz and calculate the mel-spectrograms with 64 mel-bins at two scales with hop sizes of 0.0415 seconds and 0.0083 seconds, denoted as the coarse-scale and the fine-scale respectively. The window size of the two scales for short-time Fourier transform is 0.064 seconds. Then the resulting 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.

As shown in Fig.2, each convolutional network in the block has 64 filters with 3*3 size. The convolution network out of block has 128 filters with 3*3 size. The pooling size is 1*2 behind the block and 1*4 after the single convolution layer. One bidirectional gated recurrent neural network with 64 units is used.

In the training phase, we apply the binary cross-entropy loss between the predicted probability and ground truth of an audio recording. Adam[23] 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 is better than single scale CRNN with F1-value of 85.5%. We also fuse the three models by soft voting. The fusion model achieves the best score.

Table 1. Comparison of multi-scale CRNN, single-scale CRNN, CNN based model and fusion model on development set of DCASE2018 task4.

Models	Precision(%)	Recall(%)	F1-value(%)
CNN based model[9]	85.1	85.1	85.1
Single-scale CRNN[17]	83.2	80.9	82.0
Multi-scale CRNN	83.5	87.6	85.5
Fusion	87.7	89.8	88.7

3.2.2 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 show 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 / 20% of the events length collar on offsets.

Table 2. Comparison of F1-value and the error rate of multi-scale CRNN using and not using classification results for correction on development set of DCASE2018 task4.

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

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

informational 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.

Table 3. Comparison of multi-scale CRNN, single-scale CRNN, and baseline on development set of DCASE2018 task4.

Models	F1-value(%)	Error rate
Baseline[1]	14.1	1.54
Single-scale CRNN[17]	24.4	1.24
Multi-scale CRNN	29.2	1.40

4. CONCLUSIONS

In this paper, we propose multi-scale convolutional re-current 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 dependencies. 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 correct the frame-level prediction errors with classification results.

We also tried several methods to improve the system using unlabeled data but we are not satisfied with the results achieved. To further improve the system, 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.

5. REFERENCES

[1] R. Serizel, N. Turpault, H. Eghbal-Zadeh, A. P. Shah . “Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments ”. Submitted to DCASE2018 Workshop, 2018.

[2] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah. Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments. Submitted to DCASE2018 Workshop, July 2018.

[3] Kumar, A., & Raj, B. (2016, October). Audio event detection using weakly labeled data. In Proceedings of the 2016 ACM on Multimedia Conference (pp. 1038-1047). ACM.

[4] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 776-780). IEEE.

[5] Grégoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, Axel Roebel, "A Morphological Model for Simulating Acoustic Scenes and Its Application to Sound Event Detection", Audio Speech and Language Processing IEEE/ACM Transactions on, vol. 24, no. 10, pp. 1854-1864, 2016.

[6] Heittola, T., Mesaros, A., Eronen, A., & Virtanen, T. (2013). Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1), 1.

[7] Dikmen, O., & Mesaros, A. (2013, October). 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 (pp. 1-4). IEEE.

[8] Su T W, Liu J Y, Yang Y H. Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017:791-795.

[9] Lee D, Lee S, Han Y, et al. ENSEMBLE OF CONVOLUTIONAL NEURAL NETWORKS FOR WEAKLY-SUPERVISED SOUND EVENT DETECTION USING MULTIPLE SCALE INPUT[C]// DCASE 2017 Workshop. 2017.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[11] Su, T. W., Liu, J. Y., & Yang, Y. H. (2017, March). 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) (pp. 791-795).

[12] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[14] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M. (2017, March). CNN architectures for large-scale audio classification. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 131-135). IEEE.

[15] Eghbal-Zadeh, H., Lehner, B., Dorfer, M., & Widmer, G. (2016). CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE).

[16] Rui Lu, Zhiyao Duan, Changshui Zhang. Multi-scale recurrent neural network for sound event detection[R]. Technical report, DCASE2017 Challenge (September 2017), 2017.

[17] Xu, Y., Kong, Q., Wang, W., & Plumbley, M. D. (2017). Large-scale weakly supervised audio classification using gated convolutional neural network. arXiv preprint arXiv:1710.00343

[18] Salamon J, McFee B, Li P, et al. DCASE 2017 submission: Multiple instance learning for sound event detection[R]. Technical report, DCASE2017 Challenge (September 2017), 2017.

[19] Chou, S. Y., Jang, J. S. R., & Yang, Y. H. (2017). FrameCNN: a weakly-supervised learning framework for frame-wise acoustic event detection and classification. Recall, 14, 55-4.

[20] Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In Advances in neural information processing systems (pp. 2643-2651).

[21] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M. (2017, March). CNN architectures for large-scale audio classification. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 131-135). IEEE.

[22] Eghbal-Zadeh, H., Lehner, B., Dorfer, M., & Widmer, G. (2016). CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE).

[23] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.