LARGE-SCALE WEAKLY SUPERVISED AUDIO CLASSIFICATION USING GATED CONVOLUTIONAL NEURAL NETWORK

Yong Xu, Qiuqiang Kong*, Wenwu Wang, Mark D. Plumbley*

Center for Vision, Speech and Signal Processing, University of Surrey, UK
{yong.xu, q.kong, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

In this paper, we present a gated convolutional neural network and a temporal attention-based localization method for audio classification, which won the 1st place in the large-scale weakly supervised sound event detection task of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 challenge. The audio clips in this task, which are extracted from YouTube videos, are manually labelled with one or more audio tags, but without time stamps of the audio events, hence referred to as weakly labelled data. Two subtasks are defined in this challenge including audio tagging and sound event detection using this weakly labelled data. We propose a convolutional recurrent neural network (CRNN) with learnable gated linear units (GLUs) non-linearity applied on the log Mel spectrogram. In addition, we propose a temporal attention method along the frames to predict the locations of each audio event in a chunk from the weakly labelled data. The performances of our systems were ranked the 1st and the 2nd as a team in these two sub-tasks of DCASE 2017 challenge with F value 55.6% and Equal error 0.73, respectively.

Index Terms— DCASE2017 challenge, weakly supervised sound event detection, audio tagging, attention, gated linear unit

1. INTRODUCTION

Audio classification is a task to classify audio recordings into different classes. Weakly labelled audio data contains only the presence or absence of the audio events but without the time stamps of the audio events [1]. Weakly labelled audio classification has many applications in information retrieval [2], surveillance of abnormal sound in public area and industry use [3]. Some challenges divide audio classification into subtasks including audio scene classification [4] and sound event detection [4]. Recently a large-scale weakly supervised sound event detection task was proposed as a part of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 challenge [5]. In this challenge, the data set is a subset of Google Audio Set [6] containing both transportation and warning sounds. This task includes an audio tagging (AT) [7] subtask and a weakly supervised sound event detection (SED) [8] subtask. The AT task aims to predict one or several labels of an audio recording and SED needs to predict the time stamps of the audio events.

Many audio classification methods are based on the bag of frames [9] assumption, where an audio recording is cut into segments and each segment inherits the labels of the audio recording. However this assumption is incorrect because some audio events only happen for a short time in an audio clip. Multi-instance learning (MIL) [1] has been applied to train on weakly labelled data. Recently state-of-the-art audio classification methods [10, 11] transform the waveform to the time-frequency (T-F) representation. Then the T-F representation is treated as an image which is fed into CNNs. However, unlike image classification where the objects are usually centered and occupy a dominant part of the image, audio events may only occur in a short part in an audio recording. To solve this problem, some attention models [12] for audio classification are applied to attend to the audio events and ignore the irrelevant features.

In this paper, we propose a unified neural network model which fits for both the audio tagging task and the weakly labelled sound event detection task, simultaneously. The first contribution of this paper is to apply the learnable gated linear unit (GLU) [13] to replace the ReLU activation [14] after each layer of the convolutional neural network for audio classification. This learnable gate is able to control the information flow to the next layer. When a gate value is near to 1, the corresponding T-F unit is attended. When a gate value is close to 0, then the corresponding T-F unit is ignored. Following the convolutional layers, the recurrent layer is used to exploit the temporal information. Then a temporal attention method is proposed to localize the audio events in a chunk. This attention part helps to capture the audio events and ignore unrelated audio segments hence it is able to detect sound events from weakly labeled data.

The paper is organized as follows. Section 2 introduces the gated linear units in the neural network. Section 3 proposed the localization method for audio events from the weakly labeled data. Section 4 shows experiments. Section 5 summarizes and suggests the future work.

* These first two authors contributed equally to this work.

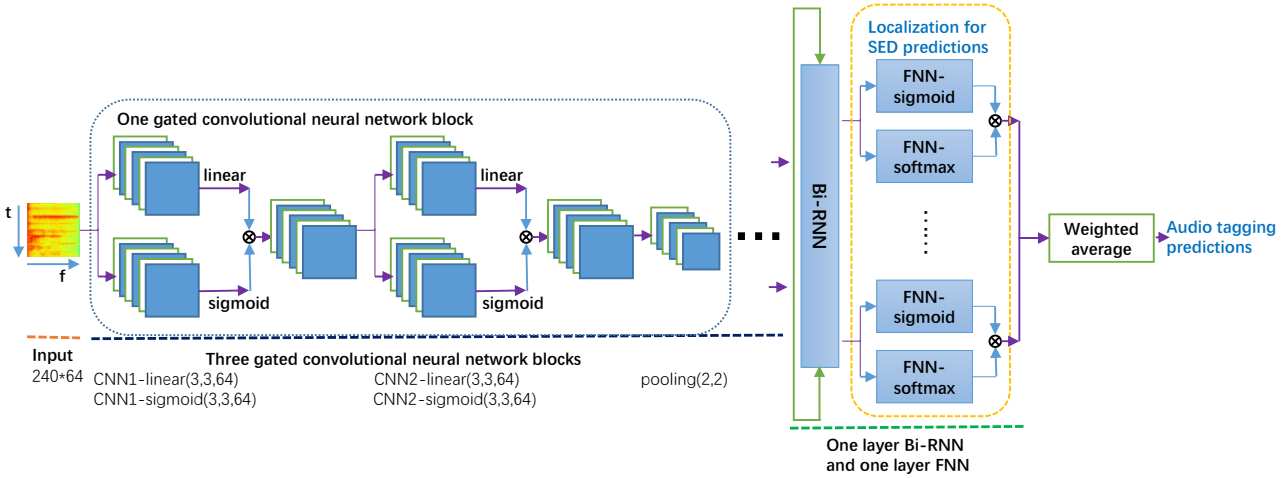

Fig. 1. The proposed unified model for audio tagging (AT) and weakly supervised sound event detection (SED). The final outputs are the AT results. SED predictions are extracted from the intermediate localization module. One gated convolutional block is shown in the dashed rectangle and the whole system has three similar blocks in total.

2. PROPOSED GATED LINEAR UNITS IN CRNN FOR AUDIO TAGGING

In this section, the convolutional recurrent neural networks (CRNNs), mini-batch data balancing, gated linear unit (GLU), and system fusion will be introduced.

2.1. CRNN

CRNNs have been successfully used in audio classification tasks [15, 11]. For audio tagging, a CRNN-based method has been proposed in [16, 12] to predict the audio tags. First the waveforms of the audio recordings are transformed to T-F representations such as log Mel spectrogram. Then convolutional layers are applied to the T-F representations to extract high level features. Then a bi-directional recurrent neural network (Bi-RNN) is adopted to capture the temporal context information, and is followed by a feed-forward neural network (FNN) to predict the posteriors of each audio class at each frame. Finally, the predicted probability of each audio tag is obtained by averaging the posteriors of all the frames.

In the training phase, we apply binary cross-entropy loss between the predicted probability and the ground truth of an audio recording. The weights of the neural network can be updated by the gradient of the loss function computed using back-propagation. The loss can be defined as:

$$E = -\sum_{n=1}^{N}(\mathbf{P}_n \log \mathbf{O}_n + (1-\mathbf{P}_n)\log(1-\mathbf{O}_n)) \quad (1)$$

where E is the binary cross-entropy, $\mathbf{O}_n$ and $\mathbf{P}_n$ denote the estimated and reference tag probability vector of the n-th audio clip, respectively. The number N represents the mini-batch size.

2.2. Mini-batch data balancing

The data set defined in this challenge is highly unbalanced, such that the number of samples of each class varies significantly. For example, the 'car' class occurred 25744 times in the data set while 'car alarm' only occurred 273 times. This highly unbalanced data will bias the training to the class with a large number of occurrences. As we are using mini-batch to train the network, there is an extreme situation where all the samples in a mini-batch are 'car'. To solve this problem we balance the frequency of different classes in a mini-batch to ensure that the number of most frequent samples is, on average, at most 5 times than the least frequent samples in a mini-batch.

2.3. Gated linear units in CNNs

We propose to use gated linear units (GLUs) [13] as activation functions to replace the conventional ReLU [14] activation functions in the CRNN model. GLUs were first proposed in [13] for language modeling. The motivation of using GLUs in audio classification is to introduce the attention mechanism to all the layers of the neural network. The GLUs can control the amount of information of a T-F unit flow to the next layer. If a GLU gate value is close to 1, then the corresponding T-F unit is attended. If a GLU gate value is near to 0, then the corresponding T-F unit is ignored. By this means the network can learn to attend to audio events and ignore the unrelated sounds. GLUs are defined as:

$$\mathbf{Y} = (\mathbf{W} * \mathbf{X} + \mathbf{b}) \odot \sigma(\mathbf{V} * \mathbf{X} + \mathbf{c}) \quad (2)$$

where σ is a sigmoid function and $\odot$ is the element-wise product and $*$ is the convolution operator. $\mathbf{W}$ and $\mathbf{V}$ are convolutional filters, $\mathbf{b}$ and $\mathbf{c}$ are biases. $\mathbf{X}$ denotes the input T-F rep-

resentation in the first layer or the feature maps of the interval layers.

The framework of the model is shown in Fig. 1. A pair of convolutional networks are used to generate the gating outputs and the linear outputs. These GLUs can reduce the gradient vanishing problem for deep networks [13] by providing a linear path for the gradients propagation while keeping non-linear capabilities through the sigmoid operation. The output of each layer is a linear projection $(\mathbf{W} * \mathbf{X} + \mathbf{b})$ modulated by the gates $\sigma(\mathbf{V} * \mathbf{X} + \mathbf{c})$. Similar to the gating mechanisms in long short-term memories (LSTMs) [17] or gated recurrent units (GRUs) [18], these gates multiply each element of the matrix $(\mathbf{W} * \mathbf{X} + \mathbf{b})$ and control the information passed on in the hierarchy [13]. From the feature selection view, the GLUs can be regarded as an attention scheme on the time-frequency (T-F) bin of each feature map. This scheme can attend to the T-F bin with related audio events by setting its value close to one otherwise to zero.

2.4. Fusion of system results

Fusion of system results is empirically important to improve the robustness of systems. In this work, we adopt two-level fusion strategies. As neural networks are trained by the gradient based optimization algorithm with a fixed or dynamically changing learning rate, the performance will be gradually better but fluctuant along the epochs. Hence, our first fusion strategy is conducted among the epochs in the same system. This will improve its stability of the system. The second fusion strategy is to average the posteriors from different systems with different configurations.

3. PROPOSED LOCALIZATION FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Different from the audio tagging task without requiring to predict the temporal locations of each audio event which is presented in Sec. 2, the sound event detection (SED) task needs to predict the temporal locations of each occurring audio event. The problem would be more difficult if there were no strong labels, namely frame-level labels. This is the so-called weakly supervised SED defined in the task 4 of DCASE2017 challenge.

As shown in the localization module of Fig. 1, an additional feed-forward neural network with softmax as the activation function is introduced to help to infer the temporal locations of each occurring class. To keep the time resolution of the input whole audio spectrogram, we adjust the pooling steps in the CNNs shown in Fig. 1 by only pooling on the spectral axis while not pooling on the time axis. So the feed-forward network with sigmoid as the activation function shown in Fig. 1 will perform classification at each frame, meanwhile the feed-forward with softmax as the activation function shown in Fig. 1 will attend to the most salient frames for each class.

If we define the FNN-softmax output $\mathbf{Z}_{\text{loc}}(t) \in \mathbb{R}^K$ as the localization vector where K is the number of output classes, then it is multiplied with the classification output $\mathbf{O}(t) \in \mathbb{R}^K$ at each frame to obtain $\mathbf{O}'(t) \in \mathbb{R}^K$:

$$\mathbf{O}'(t) = \mathbf{O}(t) \odot \mathbf{Z}_{\text{loc}}(t) \tag{3}$$

where $\odot$ represents element-wise multiplication. To obtain the final acoustic event tag predictions, $\mathbf{O}'(t)$ should be averaged across the time axis in an audio clip to obtain the final output $\mathbf{O}'' \in \mathbb{R}^K$, which is defined as the weighted average of $\mathbf{O}'(t)$ as following,

$$\mathbf{O}'' = \frac{\sum_{t=0}^{T-1} \mathbf{O}'(t)}{\sum_{t=0}^{T-1} \mathbf{Z}_{\text{loc}}(t)} \tag{4}$$

where T is final frame number and the division is element wise division. The back-propagation loss is the same as in the audio tagging task by comparing the reference labels with the final output $\mathbf{O}''$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

The task4 of DCASE2017 challenge employs a subset of Google Audio Set [6]. Audio Set consists of an large ontology of 632 sound event classes and a collection of 2 million human-labeled sound clips (mostly 10-second length) drawn from 2 million YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds. The subset used in the task consists of 17 sound events divided into two categories: "Warning" and "Vehicle".

Log Mel filter banks and Mel frequency cepstral coefficients (MFCCs) are used as features in our system. Each audio recording feature has 240 frames by 64 mel frequency channels. As shown in Fig. 1, three gated convolutional neural network blocks are adopted. Each convolutional network has 64 filters with 3*3 size. The pooling size is 2*2 for the audio tagging sub-task while the pooling size 1*2 for the sound event detection sub-task, which means no pooling is applied along time axis to maintain the time resolution for sound event detection. One bi-directional gated recurrent neural network with 128 units is used. The feed-forward neural network has 17 output nodes where each of them is corresponding to an audio event class. Adam [19] optimizer is applied and the learning rate is fixed to be 0.001. These hyper-parameters are selected empirically.

The source codes for this paper can be downloaded from our webpage on Github[1].

[1] https://github.com/yongxuUSTC/dcase2017_task4_cvssp

4.2. Results

In this section, the audio tagging results and the weakly supervised sound event detection results will be given.

4.2.1. Audio tagging

Table 1 presents the F value (F1), precision and recall [5] comparisons for the audio tagging sub-task on the development set and the evaluation set. "CRNN-logMel-noBatchBal" denotes the CRNN system trained without mini-batch data balancing strategy. The DCASE2017 baseline model was a multilayer perceptron (MLP) based method [5]. Our proposed CRNN systems show much better performance. Comparing the CRNNs with and without mini-batch balancing, we see that data balancing is important to get higher all of F1, precision and recall scores. The proposed gated CRNN also gains effective improvement. The final fusion system combines the systems trained on different features, namely log Mel and MFCC. On the evaluation set which is a blind test, our system ranks 1st in this audio tagging challenge according to the more comprehensive F1 score. Our CNN-ensemble [20] and Frame-CNN [21] ranks 2nd and 3rd, respectively. Note that our final fusion system has a notable absolute 3% improvement over the 2nd system [20].

Table 1. F1, Precision and Recall comparisons for the audio tagging sub-task on the development the evaluation sets.

Dev-set	**F1**	**Precision**	**Recall**
DCASE2017 Baseline [5]	10.9	7.8	17.5
CRNN-logMel-noBatchBal	42.0	47.1	38.0
CRNN-logMel	52.8	49.9	56.1
Gated-CRNN-logMel (i)	56.7	53.8	**60.1**
Gated-CRNN-MFCC (ii)	52.1	51.7	52.5
Fusion (i+ii)	**57.7**	**56.5**	58.9
Eval-set	**F1**	**Precision**	**Recall**
DCASE2017 Baseline [5]	18.2	15.0	23.1
CNN-ensemble [20]	52.6	**69.7**	42.3
Frame-CNN [21]	49.0	53.8	45.0
Our gated-CRNN-logMel	54.2	58.9	50.2
Our fusion system	**55.6**	61.4	**50.8**

4.2.2. Weakly supervised sound event detection (SED)

The results of F1 and Error rate comparisons on the development set and the evaluation set for the 2nd SED task are given in Table 2. Our proposed gated-CRNN-logMel method outperforms the DCASE2017 baseline [5]. With the fusion system, we rank 2nd as a team in the sound event detection sub-task. The 1st place team achieves 0.66 Error rate and 55.5% F1 score [20]. However, [20] used independent one frame input for SED, and it assumed that audio events occurred everywhere along the chunk. Our method is a unified method without any assumption. Attention based localization seems to be more reasonable for weakly supervised SED.

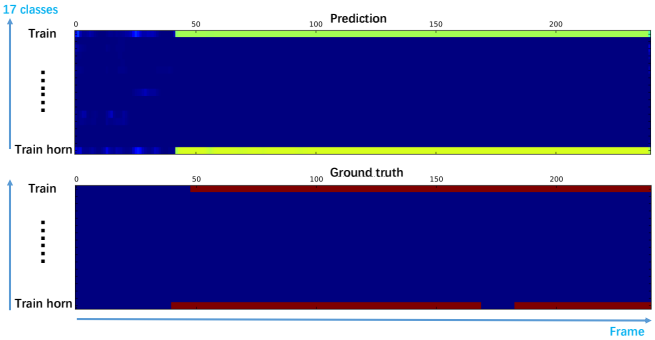

Fig. 2. An example for predicting locations along 240 frames for "10i60V1RZkQ_210.000_220.000.wav" using the proposed localization method.

Fig. 2 shows an example for predicating temporal locations along 240 frames for occurring audio events, namely 'train' and 'train horn'. Our proposed localization method can almost successfully detect the accurate temporal locations for occurring events, except for the small segment false alarm for the 'train horn'.

Table 2. The results of F1 and Error rate comparisons on the development set and the evaluation set for the **sound event detection** sub-task among several methods across the 17 audio event tags.

Dev-set	**F1**	**Error rate**
DCASE2017 baseline [5]	13.8	1.02
Gated-CRNN-logMel	47.20	0.76
Fusion	**49.7**	**0.72**
Eval-set	**F1**	**Error rate**
DCASE2017 baseline [5]	28.4	0.93
Gated-CRNN-logMel	47.50	0.78
Fusion	**51.8**	**0.73**

5. CONCLUSIONS

In this paper, we proposed a unified method for audio tagging and weakly supervised sound event detection. A gated CRNN method is proposed, where the learnable gated linear units can help to select the most related features corresponding to the final labels. A temporal attention based localization method is also proposed to localize the occurred events along the chunk in a weakly supervised mode. The final system puts us in the 1st place with 57.7% F1 score on the audio tagging sub-task of DCASE2017 challenge. We also rank 2nd as a team in the weakly supervised sound event detection sub-task. In the near future, we will evaluate our proposed gating and attention methods on Google Audio Set [6].

6. REFERENCES

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

[2] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, X. Serra, et al., "Essentia: An audio analysis library for music information retrieval.," in *ISMIR*, 2013, pp. 493–498.

[3] S. Dimitrov, J. Britz, B. Brandherm, and J. Frey, "Analyzing sounds of home environment for device recognition.," in *AmI*. Springer, 2014, pp. 1–16.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *EUSIPCO*. IEEE, 2016, pp. 1128–1132.

[5] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of DCASE2017 Workshop*.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[7] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," in *IEEE/ACM Transcation on Audio, Speech and Language Processing*, 2017.

[8] Q. Kong, Y. Xu, W. Wang, and Mark D. P. Plumbley, "A joint detection-classification model for audio tagging of weakly labelled data," in *ICASSP*, 2017, pp. 641–645.

[9] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, "Acoustic scene classification based on sound textures and events," in *Proceedings of ACM on Multimedia Conference*. ACM, 2015, pp. 1291–1294.

[10] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[11] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[12] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," in *INTERSPEECH*. IEEE, 2017, pp. 3083–3087.

[13] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[14] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[15] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," *arXiv preprint arXiv:1706.02291*, 2017.

[16] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *IJCNN*, 2017, pp. 3461–3466.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

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

[20] K. Lee, D. Lee, S. Lee, and Y. Han, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," Tech. Rep., DCASE2017 Challenge, September 2017.

[21] S. Chou, J. Jang, and Yi-H. Yang, "FrameCNN: a weakly-supervised learning framework for frame-wise acoustic event detection and classification," Tech. Rep., DCASE2017 Challenge, 2017.