A REGION BASED ATTENTION METHOD FOR WEAKLY SUPERVISED SOUND EVENT DETECTION AND CLASSIFICATION

Jie Yan[1] *Yan Song*[1] *Wu Guo*[1] *Li-Rong Dai*[1] *Ian McLoughlin*[2] *Liang Chen*[3]

[1] National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China.
[2] School of Computing, University of Kent, Medway, UK.
[3] Anhui Science and Technology Research Institute

ABSTRACT

Recently, an attention based convolutional recurrent neural network (CRNN) with learnable gated linear units (GLUs) has achieved state-of-the-art performance for audio tagging (AT) and sound event detection (SED) tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges. The introduction of GLU and temporal attention-based localization mechanisms plays an important role for both AT and SED tasks. In this paper, we propose a novel region based attention method to further boost the representation power of the existing GLU based CRNN. Specifically, we insert a feature selection (FS) structure after each GLU to create what we term a *GLU-FS* block, to exploit channel relationships. Furthermore, we extract region features (or the prototypes of certain sound events) from multi-scale sliding windows over higher convolutional layers, which are fed into an attention-based recurrent neural network to model their context information for AT and SED tasks. To evaluate the proposed region based attention method, we conduct extensive experiments on SED and AT tasks in DCASE2017. We achieve 59.5% and 60.1% AT F1-score, 51.3% and 55.1% SED F1-score for development and evaluation sets respectively, significantly outperforming state-of-the-art results.

***Index Terms*—** sound event detection, audio tagging, weakly labelled data, attention

1. INTRODUCTION

Sound event detection (SED) is the task of determining not only event class but also the time boundaries of events occurring in continuous audio. SED has attracted increasing research interest since auditory information can be more useful than visual information in tasks such as monitoring fire alarms, detecting gunshots in public areas or responding to cries of human distress [1]. Furthermore, it is beneficial to exploit both auditory as well as visual information in many applications including video retrieval, surveillance, and healthcare [2, 3, 4].

However, real-life SED is challenging since multiple sound events can overlap, and may exhibit many long or short-term dependencies. In addition, it is difficult to collect large-scale datasets with strong labels that provide reliable temporal information about each sound event. Recently, a large scale weakly supervised SED task was conducted in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, consisting of two closely related sub-tasks: Audio Tagging (AT) and SED, given audio training data excerpted from the large-scale Google Audio Set, tagged to indicate the presence or absence of sound events [5].

It is important to find the effective representation for both AT and SED tasks. In DCASE, the baseline is implemented using a multi-layer perceptron architecture with log mel-band energies as input [5]. Alternatives found in the literature include systems based on CNN, RNN and CRNN structures [6, 7, 8, 9]. For example Hershey et al. [6] evaluated different CNN architectures such as VGG [10], Inception [11] and ResNet [12] for large-scale audio classification. Lu et al. [7] proposed a multi-scale RNN model to capture long-term dependency by integrating information from different time resolutions. Xu et al. [8] introduced a deep convolutional recurrent model to predict both audio tags and temporal locations. In that system, the attention mechanism identifies the importance of each audio frame. The same authors then presented [9] a CRNN with learnable GLU non-linearity [13] to exploit the attention mechanism for time-frequency (T-F) units in convolutional layers. This structure, shown in Fig. 1, demonstrated excellent performance for AT and SED tasks in DCASE2017.

This paper proposes the novel region based attention method, shown in Fig. 2, to boost the representational power of the GLU based CRNN architecture. It contributes two main areas of improvement. (a) First, the GLU-FS block, a GLU with additional feature selection (FS) structure, is proposed to integrate both local and global attention mechanisms. GLU blocks [9] use a learnable gate is to replace

This work was partially funded by the National Natural Science Foundation of China (Grant No. U1613211, U1836219).

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

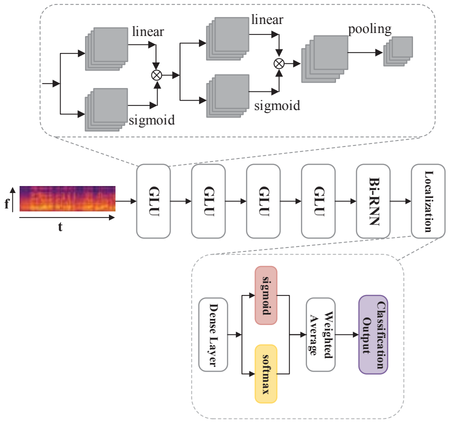

Fig. 1. Illustration of GLU based CRNN structure [9].

the ReLU [14] after each convolutional layer to control information flow to the subsequent layer. The learned gate value provides a local attention mechanism, indicating the importance of individual T-F units. Our additional FS structure in GLU-FS allows the network to perform further feature recalibration by using global T-F information to provide channel-wise selective emphasis. (b) Second, a region based attention method is proposed to model the temporal and frequency context information. This extracts region features from multi-scale sliding windows over higher convolutional layers, mainly motivated by recent advances in object detection methods like feature pyramid networks (FPN) [15]. These region-based features can effectively model T-F context information to form the prototypes of certain sound events, similar to senones in speech recognition. An attention-based recurrent neural network is then applied to model context information among these prototypes. The proposed region based attention method is evaluated through extensive experiments on DCASE2017 SED and AT tasks. It achieves 59.5% and 60.1% AT F1-score, 51.3% and 55.1% SED F1-score on development and evaluation sets respectively, significantly outperforming state-of-the-art GLU based CRNNs.

2. PROPOSED METHOD

In this section, we introduce the proposed GLU-FS structure and region based attention method, shown in Fig.2. The GLU-FS block exploits both local and global information to recalibrate the output of convolutional layers for better representation, while the region based attention method aims to model temporal and frequency context information for both AT and SET tasks.

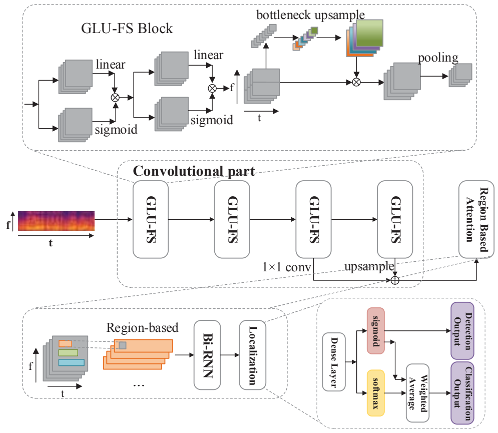

Fig. 2. Illustraion of the proposed region based attention based method, where *Detection output* refers to sound event detection, and *Classification output* refers to audio tagging.

2.1. GLU-FS Blocks

Generally, given the input feature $\mathbf{X}$, a convolutional layer can be represented as

$$\mathbf{Y} = \mathbf{U} * \mathbf{X} + \mathbf{b} \tag{1}$$

where $*$ denotes the convolutional operator, $\mathbf{U}$ and $\mathbf{b}$ are filter kernel and bias. The matrix $\mathbf{Y} \in \mathbb{R}^{C \times T \times F}$ denotes the $T \times F$ output units with C channels.

GLU blocks can be used to introduce the local attention mechanism for all convolutional layers [9]. In each GLU block, a branch that consists of a convolutional layer and sigmoid activation function is applied for learning gated units. Then the block output is the weighted version of matrix $\mathbf{Y}$;

$$\mathbf{V} = \mathbf{Y} \odot \sigma(\mathbf{Z}) \tag{2}$$

where $\mathbf{Y}$ and $\mathbf{Z}$ are the outputs of convolutional operators with different filters, σ is the sigmoid function and $\odot$ is the element-wise product.

Alternatively, the FS structure can be used to exploit global information to recalibrate the channel-wise filter responses (similar to a Squeeze-and-Excitation (SE) block [16]). Specifically, the vector $\mathbf{d} \in \mathbb{R}^{C}$ with channel information is first obtained by global average pooling over $\mathbf{Y}$;

$$\mathbf{d} = \frac{1}{T \times F} \sum_{i=1}^{T} \sum_{j=1}^{F} \mathbf{Y}_{ij} \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{C \times T \times F}$ is the feature map after the convolutional operator. Next, a bottleneck structure with two dense layers is inserted to model the interdependencies between channels, resulting in a scaling vector $\mathbf{s}$;

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{d})) \tag{4}$$

where δ is the ReLU function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ refer to weights of dense layers, and r is the reduction ratio of the bottleneck. The output of the FS structure can be computed by weighting $\mathbf{Y}$ with $\mathbf{s}$;

$$\mathbf{E} = \mathbf{s} \cdot \mathbf{Y}. \tag{5}$$

Compared to the GLU block, the FS structure introduces a global attention mechanism by explicitly modelling interdependencies between channels.

The proposed GLU-FS block (shown in Fig. 2) thus exploits both local and global attention mechanisms. It is worth mentioning that we divide the feature map into two parts (high and low frequency respectively) for pooling, considering the obvious distribution in the frequency axis. As a result, a more expressive representation is obtained by unit-wise and channel-wise attention.

2.2. Region Based Attention Method

For AT and SED tasks, CRNN based methods have achieved success. CRNN used a bi-directional recurrent neural network (bi-RNN) to capture temporal context information, followed by a feed-forward neural network (FNN) to predict the posteriors of each event class at each frame. Xu et al. [9] used an additional FNN with softmax activation function to help infer the temporal locations of each sound event. Actually, this FNNs provided a temporal attention mechanism based on the predicted posteriors.

However, the frame-level prediction was not accurate enough for AT and SED tasks. To address this, our proposed region based attention method extracts region features using sliding windows. Considering the characteristics of different sound events, various window sizes are used to generate region-based features as input to a bi-RNN. These region features contain rich information about sound events, performing a similar task to senones in speech recognition, and helping to improve prediction. Although the output of the final block is semantically stronger, it has lower time and frequency resolution. Therefore, inspired by FPN [15], we fuse feature maps from different layers by using upsampling and element-wise addition operation, as shown in Fig. 2.

3. IMPLEMENTATION

Model. The model shown in Fig. 2 is designed for both AT and SED tasks, which consists of four GLU-FS blocks in the convolutional part, followed by a region based attentional bi-RNN. In GLU-FS blocks, the convolution layers have 64 channels with 3×3 kernel size and a max-pooling layer of size 1×2. The size of the final GLU-FS block output feature map is 240×4 size. This is upsampled to match the $240{\times}8$ size of the previous block output and added to it. Three kinds of region are then extracted from the combined map, sized $2{\times}4$, $8{\times}4$ and $10{\times}4$ respectively. The stride in each is $2{\times}4$.

Regarding the elements in each region, we use l2 norm to obtain an output activation. These region-based features are then fed into one bi-directional gated recurrent layer with 128 cells and a dense layer with 128 neuron units respectively. The output of the dense layer with sigmoid activation is the detection output, while the weighted average of that and the dense layer softmax activation forms the classification output, all as shown in Fig. 2. During model training, we use the Adam [17] method. The input feature map is 240 frames of 64 log-Mel filter banks.

Loss Function. The loss function for training the model is a weighted sum of detection and classification losses;

$$L_{reg} = \sum_{n=1}^{N} L_{dec}(\hat{\mathbf{O}}_n, \hat{\mathbf{P}}_n) + \lambda L_{cls}(\mathbf{O}_n, \mathbf{P}_n) \tag{6}$$

where L_{dec} and L_{cls} are both binary cross-entropy and λ is a positive coefficient to trade-off between two kinds of losses. $\hat{\mathbf{O}}_n$ and $\hat{\mathbf{P}}_n$ are detection output, upsampled to frame number and detection label for sample n. $\mathbf{O}_n$ and $\mathbf{P}_n$ are classification output and classification label for n respectively. N is the size of one batch in the training step. The final loss is the weighted sum loss for different regions.

Event Activity Detection. As mentioned before, we use both detection loss on frame prediction and classification loss on clip prediction to train the model jointly. To get detection labels, Hershey et al. [6] and Serizel et al. [18] simply assumed that the label of each frame is the same as the label of the whole audio clip. However, the silent regions of each audio clip also hold sound event labels, but do not contribute to that sound. We therefore use an event activity detection (EAD) technique to improve the specificity of the training labels. This is accomplished using an energy threshold of each frame to decide if it contains a specific event or not:

$$th_k = \alpha_k * \bar{e}, \quad where \quad \bar{e} = \frac{1}{T} \sum_{t=1}^{T} e_t \tag{7}$$

where e_t is the energy of frame t and $\bar{e}$ is the average energy of the audio clip. α_k is a normalization coefficient for each sound event k (but is simply set to 0.1 for every sound event in our experiment). If the energy of a frame is lower than th_k, it is considered as silence. Otherwise, the frame inherits the label of the overall audio clip.

4. EXPERIMENTS

4.1. Data and Systems

We use the dataset of task 4 in the DCASE2017 challenge [5], which is a subset of AudioSet [19]. The dataset consists of 17 sound events divided into two categories: “Warning” and “Vehicle”, which aim at industry relevance. The data contains 51,172 training clips, 488 development clips, and 1,103 evaluation clips. All these clips are less-than 10-seconds.

Table 1. DCASE2017 audio tagging (AT) task results for proposed and state-of-the-art methods.

Development set	**F1**	**Precision**	**Recall**
GLU-CRNN	55.2	53.0	57.6
SE-CRNN	56.5	54.2	59.1
GLU-FS-CRNN	57.6	55.4	59.9
GLU-FS-RA8×2	56.1	54.1	58.3
GLU-FS-PF-RA1×4	56.5	53.9	59.2
GLU-FS-PF-RA8×4	57.6	54.4	61.2
GLU-FS-PF-RA8×8	56.3	53.3	59.7
Gated-CRNN-logMel [9]	56.7	53.8	60.1
GLU-FS-final	**59.5**	**56.1**	**63.4**
Evaluation set	**F1**	**Precision**	**Recall**
GLU-CRNN	56.1	53.3	59.2
SE-CRNN	56.5	53.6	59.8
GLU-FS-CRNN	57.4	54.3	60.9
GLU-FS-RA8×2	57.9	55.6	60.4
GLU-FS-PF-RA1×4	56.8	54.2	59.6
GLU-FS-PF-RA8×4	58.4	55.1	62.1
GLU-FS-PF-RA8×8	58.0	54.5	62.0
Gated-CRNN-logMel [9]	54.2	**58.9**	50.2
GLU-FS-final	**60.1**	56.9	**63.6**

Table 2. DCASE2017 sound event detection (SED) task results for proposed and state-of-the-art methods.

Development set	**F1**	**Error rate**
GLU-CRNN	47.7	0.73
SE-CRNN	48.6	0.75
GLU-FS-CRNN	48.6	0.73
GLU-FS-RA8×2	48.3	0.76
GLU-FS-PF-RA1×4	48.7	0.75
GLU-FS-PF-RA8×4	49.3	0.76
GLU-FS-PF-RA8×8	47.3	0.76
Gated-CRNN-logMel [9]	47.2	0.76
GLU-FS-final	**51.3**	**0.71**
Evaluation set	**F1**	**Error rate**
GLU-CRNN	50.6	0.76
SE-CRNN	52.0	0.77
GLU-FS-CRNN	51.9	0.76
GLU-FS-RA8×2	52.3	0.76
GLU-FS-PF-RA1×4	51.7	0.79
GLU-FS-PF-RA8×4	53.1	0.78
GLU-FS-PF-RA8×8	51.8	0.76
Gated-CRNN-logMel [9]	47.5	0.78
GLU-FS-final	**55.1**	**0.73**

Since our method uses log-Mel spectrograms as input, we fairly compare it with the gated-CRNN-logMel system [9].

We also compare to GLU-CRNN. the architecture proposed in [9], but using our loss in eqn. 6 in the training step. We then obtain results for variants of the system that use SE blocks (SE-CRNN) and GLU-FS blocks (GLU-FS-CRNN) respectively. Moving to our proposed method, GLU-FS-RAx×y refers to a system using both GLU-FS blocks and the proposed region-based method, with region size x×y. GLU-FS-PF-RAx×y then applies the pyramid structure prior to region selection. The final proposed configuration is referred to here as GLU-FS-final. By conducting numerous experiments on different variant systems, we are able to effectively assess the relative contribution of each of the proposed changes to final performance.

4.2. Results and Analysis

We present and analyze results for DCASE2017 sound event detection (SED) and audio tagging (AT) tasks for different variant systems in Tables 1 and 2 respectively. Evaluation and development set results are shown separately.

Sound Event Classification or AT results presented in Table 1 show that our proposed GLU-FS-final method performs better than the gated-CRNN-logMel system. Analyzing the variant F1 scores, we can see that SE improves on GLU. And adding the FS structure (i.e. GLU-FS-CRNN) achieves its aim of selecting distinctive features to improve results further. Considering different region sizes, the size of feature representation to generate the region is 240×8 (time axis by frequency axis) and we can see that a region size of 8×4 achieves better performance than either 1×4 or 8×8. This shows that sound event prototypes with a longer time duration and a shorter frequency span are merited. The GLU-FS-final system then combines three scales of region sizes (2×4, 8×4 and 10×4) to obtain multi-scale features and yield excellent F1 score.

Sound Event Detection or SED results presented in Table 2 show that our proposed GLU-FS-final method also outperforms the gated-CRNN-logMel system both in terms of F1 score and error rate. Examining the results of variant systems, we see a similar trend to the AT results for F1 score, apart from there being less relative benefit in moving from SE-CRNN and GLU-CRNN to GLU-FS-CRNN. Meanwhile the same trend of F1 score is found for different region sizes, indicating a similar trade-off between frequency and time axis as in the AT task. It is noticed that the error rate of GLU-FS-PF-RA$x \times y$ is slightly worse, which is likely due to the decrease in spatial resolution. We will address this in the future work. Again, the proposed GLU-FS-final architecture significantly outperforms all other systems, including the best reported state-of-the-art Gated-CRNN-logMel system.

5. CONCLUSION

This paper presents a novel multi-scale region based attention method for sound event detection and classification. The proposed method uses a feature selection structure at the output of each gated linear unit block, combined with region-based features, to incorporate both local and global information. Results obtained from tests on the DCASE2017 challenge audio tagging and sound event detection tasks show that the model outperforms the gated-CRNN-logMel system, especially in the sound event classification task.

6. REFERENCES

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

[2] Y. Wang, S. Rawat, and F. Metze, "Exploring audio semantic concepts for event-based video retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1360–1364.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 52:1–52:46, 2016.

[4] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[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 *Detection and Classification of Acoustic Scenes and Events 2017 Workshop(DCASE)*, 2017.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[7] R. Lu, Z. Duan, and C. Zhang, "Multi-scale recurrent neural network for sound event detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 131–135.

[8] 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.

[9] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 121–125.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[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] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

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

[18] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[19] 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.