SELF-ATTENTION MECHANISM BASED SYSTEM FOR DCASE2018 CHALLENGE TASK1 AND TASK4

Jun Wang

Beijing University of
Posts and Telecommunications
wangjun19930314@bupt.edu.cn

Shengchen Li

Beijing University of
Posts and Telecommunications
shengchen.li@bupt.edu.cn

ABSTRACT

In this technique report, we provide self-attention mechanism for the Task1 and Task 4 of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE2017) challenge. We take convolutional neural network (CNN) and gated recurrent unit (GRU) based recurrent neural network (RNN) as our basic systems in Task 1 and Task 4. In this convolutional recurrent neural network (CRNN), gated linear units (GLUs) is used for non-linearity which implement a gating mechanism over the output of the network for selecting informative local features. Self-attention mechanism called intra-attention is used for modeling relationship between different positions of a single sequence over the output of the CRNN. Attention-based pooling scheme is used for localizing the specific events in Task 4 and for obtaining the final labels in Task 1. In a summary, we get 70.81% accuracy subtask 1 of Task 1. In the subtask 2 of Task 1, we get 70.1% accuracy for device a, 59.4% accuracy for device b, and 55.6 accuracy for device c. For Task 1, we get 26.98% F1 value for sound event detection in old test data of developmemt data.

Index Terms— GCNN, GLU, attention mechanism

1. INTRODUCTION

Sounds contain a variety of information that humans use to understand the surroundings, without visual information, humans can easily recognize the scenes and events from the surrounding sounds because our auditory system is well trained.

AED is a closely related research area to ASC. An acoustic scene may be thought as a collection of sound events on top of some ambient noise. For instance, a "park" scene may be identified from bird chirping sound, a "restaurant" scene may be identified from cutlery, dishes and people's conversations sounds and a 'bus' scene may be identified from engine, braking and door opening sounds. It is difficult to create an automated system that recognize acoustic scenes and events, because it needs high level of information.

There are many applications of ASC and AED including multimedia indexing [1], intelligent monitoring system in living environment [2], scene classification and recognition [3], automatic audio tagging [4], audio segmentation [5], and health care [6], etc.

Some approaches to ASC [7] exploit binaural representation techniques to increase the scene classification accuracy. Sound event detection performance can be improved using ASC [8]. ASC and sound event detection are closely related, and the boundary between them is often blurred [9].

For weekly label sound event detection, some methods based on deep neural network have been introduced in recent years. Kumar et.al proposed a two frameworks based on multiple instance learning [10], one based on support vector machines, and the other on neural networks. Kong et.al [11] proposed a joint detection and classification (JDC) framework trained only on weakly labelled audio data. T-F segmentation framework is proposed to estimate the presence probability of each sound event and predict onset and offset times is from the T-F segmentation masks for SED [12] .

Here we limit the scope to identification of environmental sounds and detection of weekly label sound event.

2. PROPOSED ARCHITECTURE

We present a bunch of methods to solve Task 1 and Task 4, including mixup data augumention, gated activation function, self-attention mechanism and incremental learning scheme.

2.1 Features

Mel-frequency Cepstral Coefficients (MFCCs) have been used inclusivly in acoustic sound classification [13][14]. In recent works of sound event detection [15] [16], the use of MFCCs is shown that because of being sensitive to background noise it is not the best choice.
In speech recognition, Mel-filter bank (MBK) features have already been demonstrated to be better than MFCCs in the deep neural network [17]. In this report we take log-Mel filter banks as features.

In Task 1, subtask 1 uses 48KHz sampling rate, and subtask 2 uses 44.1 kHz sampling rate of, so for this task, we uniformly use 44.1 kHz sampling rate. Each 10-second chunk has 320 frames by 128 mel frequency channels.

In Task 4, the sampling rate of the audio segment is different, we uniformly use 16 kHz sampling rate. Each 10-second chunk has 240 frames by 64 mel frequency channels.

2.2 Data augumention

This report uses mixup as a method of data enhancement [18]. This method can improve the generalization ability of the model and construct a virtual training sample. The mathematical expression for the mixup is as follows:

$$\hat{x} = \lambda x_i + (1-\lambda) x_j$$
$$\hat{y} = \lambda y_i + (1-\lambda) y_j$$

Where (x_i, y_i) and (x_j, y_j) are two samples randomly extracted from the training data, and $\lambda \in [0,1]$, $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, $\alpha \in (0, \infty)$. The mixup extends the characteristics of the training set and the label distribution by linear interpolation of the feature vectors and linear interpolation of the corresponding labels. The super-parameter α of the mixup controls the interpolation strength of the features and labels

2.3 Gated linear units

In this report, our baseline system references to previous work [19]. In this baseline system, we use a learnable gated activation function called GLU [20] rather than sigmoid or ReLU to introduce the non-linear characteristics in CRNN network. GLUs are defined as:

$$Y = (W * X + B) \odot \sigma(V * X + C)$$

Where σ is the sigmoid nonlinear activation function, $\odot$ denotes the corresponding element point multiplication, and $*$ denotes the convolution operator. W and V represent the filters in the convolutional layer. In the input layer, X represents the time frequency of the input in the first layer, and in the middle layer, X represents the input of the intermediate layer.

2.3 Self-attention structure

In this report, we introduce self-attention mechanism proposed in previous work [21]. In previous work, attention mechanism is described as a kind of mapping from a Q and a set of K-V pairs to an output, where the Q, K, V, and output are all vectors. The output of attention is a weighted sum of the V, where the weight is assigned to each value which is computed by dot-product of the Q with the corresponding K. When K, V and Q come from the same source, this kind of attention is call self-attention mechanism. In the self-attention mechanism, vector of each position could process all positions in the output of previous layer. In this report, CRNN structure is followed by self-attention mechanism. Self-attention structure is shown as Fig 1.

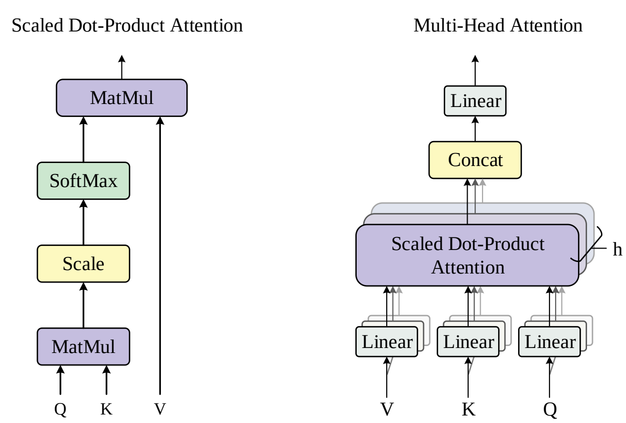


Fig 1: The diagram of the proposed self-attention

2.4 Overall neural network structure

Our system is based on the previous work [19], and the global weighted pooling (GWP) is used in the last layer rather than global maximum pool and global average pool. The block diagram of the system structure used in this paper is shown in Fig 2.

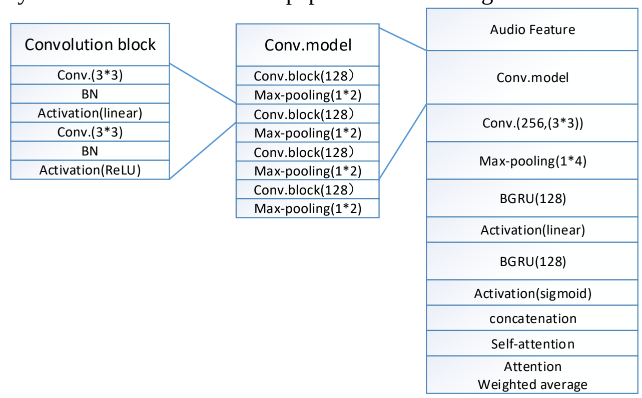


Fig 2: The diagram of the proposed unified model in Task 1 and Task 4

2.4 Incremental learning scheme

Because the Task 4 has unlabeled dada, this algorithm is only used in Task 4. At the beginning, the dataset L is a small amount of labeled data; we train a model $\mathcal{M}_0$ with a small amount of labeled date and run it on U to select b number of samples for labeling according to pre-defined thresholds α . The newly labeled samples will be incorporated into L to continuously fine-tune the pre-trained model incrementally until the number of samples in U is less than pre-defined N. Several researchers have demonstrated that fine-tuning offers better performance and is more robust than training from scratch. The incremental algorithm is illustrated in Alg. 1

Algorithm 1: Incremental fine-learning method

Input:

$U = \{C_i\}$, $i \in [1, n]$, {U contains n candidates， without label}

$\mathcal{M}_0$: pre-trained CNN

b: batch size

α: pre-defined thresholds to pick samples worthy of labeling

N: pre-defined number to stop fine tuning when the number of samples without labels is less than N

Output：

$\mathcal{L}$: Labeled samples

$\mathcal{M}_t$: fine-tuned model at Iteration t

Functions:

$p \leftarrow P(C, \mathcal{M})$ {outputs of $\mathcal{M}$ given $\forall_x \in C$}

$\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$ {fine-tune $\mathcal{M}_{t-1}$ with $\mathcal{L}$}

Initialize：

$\mathcal{L} \leftarrow$ Labeled training set, $t \leftarrow 1$

1 repeat
2 for each $C_i \in$ U do
3 $p_i \leftarrow \mathrm{P}(C_i, \mathcal{M}_{t-1})$
4 End
5 according to the category, put the samples which the predicted value p_i is larger than α into the new set Q and deduplicate the set Q
6 $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{Q}$; $\mathrm{U} \leftarrow \mathrm{U} \setminus \mathcal{Q}$
7 $\mathcal{M}_t \leftarrow \mathrm{F}(\mathcal{L}, \mathcal{M}_{t-1})$, $\mathrm{t} \leftarrow \mathrm{t}+1$
8 until the number of samples in U is less than N

3. EXPERIMENT AND RESULTS

The parameters of the network were initialized by random values sampled from zero-mean normal distribution. These systems of Task1 and Task4 are trained by using back-propagation with binary cross-entropy loss function, correct labels and estimated labels. A stochastic gradient descent algorithm [22] is performed using Adam algorithm optimization [23] in mini-batches to improve learning convergence. Dropout technique [24] is used to prevent overfitting problem.

3.1. Task 1: Acoustic Scene Classification

TUT Acoustic scenes 2018 dataset is used in this task. This task contains three subtasks, we solve subtask1 and subtask2. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. Each recording contains 10-second segments. The development data consists of recordings from all six cities.

For subtask1, of the total 8640 segments, 6122 segments are used for training, 1000 segments are used for validation and 2518 segments used for testing.

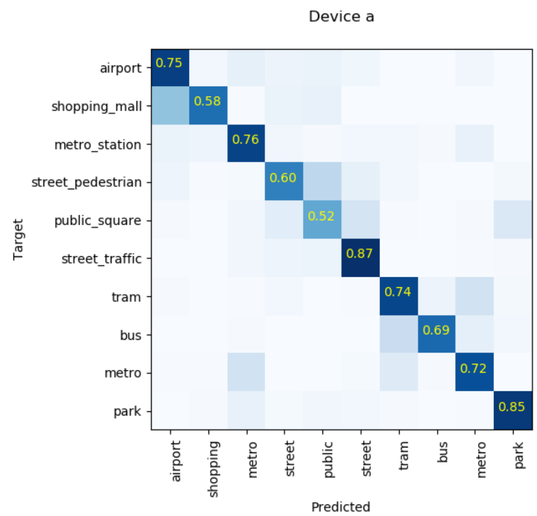
We compare the results of baseline system structure (without self-attention) with the proposed structure with self-attention. The results are shown in Table 1.

Table 1: Accuracy of subtask 1 in Task 1

Scene label	Baseline	Self-attention
airport	68.30	75.47
shopping mall	59.86	58.06
metro station	80.69	76.06
street pedestrian	67.21	59.51
public square	46.30	52.31
street traffic	87.40	86.59
tram	81.61	74.33
bus	62.81	69.01
metro	69.35	72.03
park	83.06	84.71
Average	70.66	70.81

Confusion matrix of the result of subtask 1 with self-attention is shown in Fig 3.

Fig 3: Confusion matrix of the result of subtask 1 with self-attention



For subtask 2, of the total 10080 segments, 6202 segments are used for training, 1000 segments are used for validation and 2878 segments are used for testing.

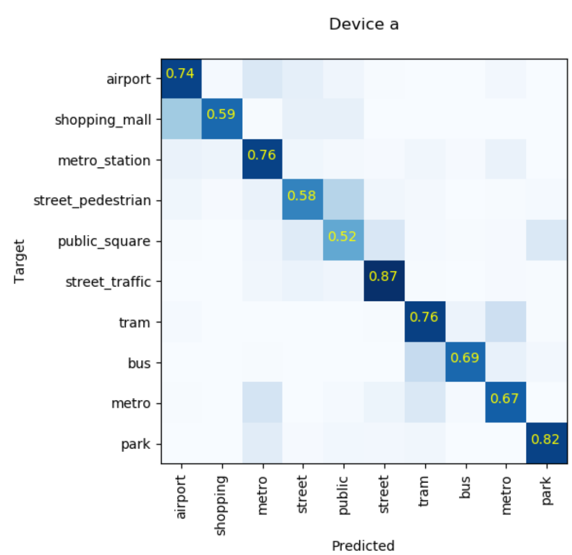
We compare the results of baseline system structure (without self-attention) with the proposed structure with self-attention. The overall average results for three devices are shown in Table 2.

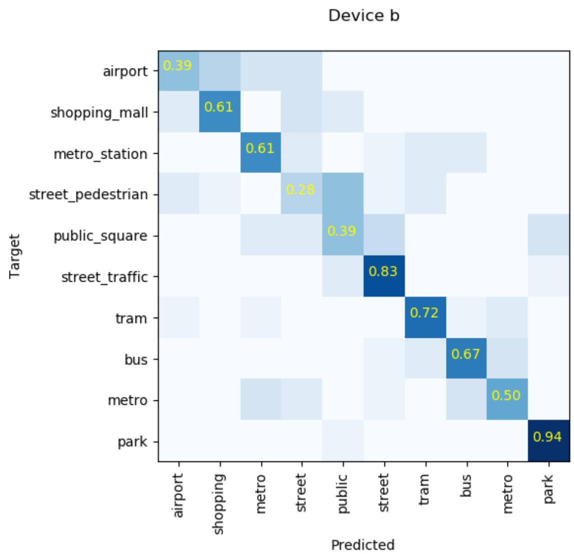
Table 1: Overall accuracy of subtask 2 in Task 1

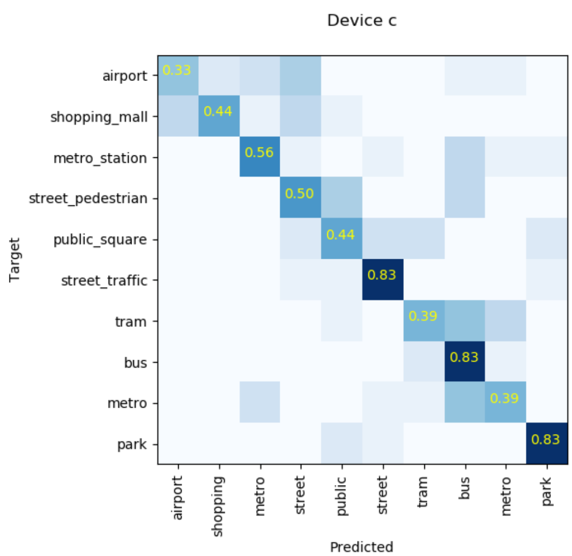
Device	Baseline (%)	Self-attention (%)
a	61.9	70.1
b	49.4	59.4
c	48.9	55.6
Average b/c	49.2	57.5

The confusion matrix of three devices of subtask 2 are shown in Figure 4 .

Fig 4: The confusion matrix of three devices of subtask 2 with self-attention







3.1. Task 4: Large-scale weakly labeled semi-supervised sound event detection in domestic environment

The data from You tube video which excerpts from domestic context are used in Task 4.

The objective of this task is to provide not only the event class but also the event time boundaries given that multiple events can be present in a 10-second audio chunk. There are 10 kinds of events occurring in audio segments including Speech, Dog, Cat, Alarm/bell/ringing, Dishes, Frying, Blender, Running water, Vacuum cleaner and Electric shaver/toothbrush.

This task provides 3 different splits of training data and one test dataset with strong label. Training dataset contains three parts: labeled training set, unlabeled in domain training set and unlabeled out of domain training set. Labeled training set contains 1578 clips (2244 class occurrences) with weak annotations, while unlabeled in domain training set with 14412 clips and unlabeled out of domain training set with 39999 clips have no labels. Because the distribution of Audio labels on unlabeled out of domain training set might not be similar with labeled training set, unlabeled in domain training set, we will discard unlabeled out of domain training set in the experiment.

We compare the results of baseline system structure with the proposed structure based on self-attention. The results are shown in Table 3.

Table 3: F1, Precision and Recall comparisons for the on the development datasets

Model	F1 (%)	P (%)	R (%)
Baseline	23.67	23.00	24.39
Baseline(Incremental)	26.98	28.79	25.39
Self-attention	8.01	10.12	6.62

Table 4: Class-wise performance comparisons of baseline and self-attention

Event label	Baseline			Self-attention		
	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)
Speech	47.9	55.9	41.9	0.0	0.0	0.0
Dog	1.8	2.0	1.6	2.2	3.8	1.6
Cat	2.8	2.3	3.7	3.1	4.4	2.4
Alarm bell-ing	31.5	40.3	25.9	3.8	6.2	2.7
Dishes	12.0	18.0	9.0	1.2	2.0	0.8
Frying	15.0	9.6	33.3	30.8	24.4	41.7
Blender	18.0	16.3	20.0	16.1	14.9	17.5
Running-water	28.8	21.0	46.1	8.0	7.1	9.2
Vacuum clean	12.7	9.5	19.4	42.9	37.5	50.0
Electric shaver	30.7	33.3	32.1	31.3	27.8	35.7

4. CONCLUSIONS

This technique report briefly describes the overall framework and some methods for the Task 1 and Task 4 of DCASE2018 challenge. We found the self-attention mechanism can improve the performance of the system effectively in Task 1. In Task 4, the incremental learning algorithm can effectively improve the performance of the system. Although the self-attention mechanism does not improve the overall performance of the system, it improves the performance of some sound events, for example, “frying”, Vacuum clean and “Electric shaver”, which is long duration.

5. REFERENCES

[1] D. Zhang and D. Ellis, “Detecting sound events in basketball video archive,” *Dept. Electron. Eng. Columbia Univ*, 2001.

[2] D. Stowell and D. Clayton, “Acoustic event detection for multiple overlapping similar sources,” *2015 IEEE Work. Appl. Signal Process. to Audio Acoust. WASPAA 2015*.

[3] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[4] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, “Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices,” *2012 IEEE Int. Conf. Emerg. Signal Process. Appl. ESPA 2012 - Proc.*, 2012.

[5] G. Wichern *et al.*, “Segmentation , Indexing , and Retrieval for Environmental and Natural Sounds,” vol. 18, no. 3, 2010.

[6] J. Schroeder, S. Wabnik, P. W. J. van Hengel, and S. Goetze, “Detection and Classification of Acoustic Events for In-Home Care,” *Ambient Assist. Living*, 2011.

[7] Y. Han, J. Park, and K. Lee, “Convolution Neural Networks with Binaural Representation and Background Subtraction for Acoustic Scene Classification,” vol. 2, no. November, 2017.

[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” pp. 1–13, 2013.

[9] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic Scene Classification: Classifying environments from the sounds they produce,” *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, 2015.

[10] A. Kumar and B. Raj, “Audio Event Detection using Weakly Labeled Data,” 2016.

[11] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, “A joint separation-classification model for sound event detection of weakly labelled data,” 2017.

[12] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, “Sound Event Detection and Time-Frequency Segmentation from Weakly Labelled Data,” pp. 1–10, 2018.

[13] Y. Han and K. Lee, “Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation,” no. July, 2016.

[14] R. Cai, L. Lu, A. Hanjalic, and H. Zhang, “A Flexible Framework for Key Audio Effects Detection and Auditory Context Inference,” vol. 14, no. 3, 2006.

[15] P. Hamel and D. Eck, “Learning features from music audio with deep belief networks,” in *ISMIR*, 2010, pp. 339–344.

[16] C. V. Cotton and D. P. W. Ellis, “Spectral vs. spectro-temporal features for acoustic event detection,” *IEEE Work. Appl. Signal Process. to Audio Acoust.*, pp. 69–72, 2011.

[17] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 7398–7402, 2013.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” pp. 1–13, 2017.

[19] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” pp. 2–6, 2017.

[20] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language Modeling with Gated Convolutional Networks,” 2016.

[21] A. Vaswani *et al.*, “Attention Is All You Need,” no. Nips, 2017.

[22] R. Johnson and T. Zhang, “Accelerating Stochastic Gradient Descent using Predictive Variance Reduction,” *Nips*, vol. 1, no. 3, 2013.

[23] A. T. Hadgu, A. Nigam, and E. Diaz-Aviles, “Large-scale learning with AdaGrad on Spark,” *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, vol. 2, 2015.

[24] G. Hinton, “Dropout : A Simple Way to Prevent Neural Networks from Overfitting,” vol. 15, 2014.