TIME-FREQUENCY SEGMENTATION ATTENTION NEURAL NETWORK FOR URBAN SOUND TAGGING

Technical Report

Lin Cui, Shaonan Ji, Xinyuan Han, Jinjia Wang,

Yanshan Univeisity
School of Information Science and Engineering, Department of Electronic communication
Qin Huangdao, Hebei, China

ABSTRACT

Audio tagging aims to assign one or more labels to the audio clip. In this task, we used the Time-Frequency Segmentation Attention Network (TFSANN) for urban sound tagging. In the training, the log mel spectrogram of the audio clip is used as input feature, and the time-frequency segmentation mask is obtained by the time-frequency segmentation network. The time-frequency segmentation mask can be used to separate the time-frequency domain sound event from the background scene, and enhance the sound event that occurred in the audio clip. Global Weighted Rank Pooling (GWRP) allows existing event categories to occupy significant part of the spectrogram, allowing the network to focus on more significant features, and it can also estimate the probability of existence of sound event. In this paper, the proposed TFSANN model is validated on the development dataset of DCASE2019 task 5. Finally, the coarse-grained and fine-grained taxonomy results are obtained on the Micro Area under precision-recall curve (AUPRC), Micro F1 score and Macro Area under precision-recall curve (AUPRC).

Index Terms— DCASE2019, Audio Tagging, Time-frequency Segmentation Network, Attention

1. INTRODUCTION

The sounds in our everyday environment carry a lot of information about events happening nearby, but the machine sound processing is still far behind. Audio tagging plays an important role in multimedia understanding. The Sound of New York City (SONYC) is a system for monitoring, analysing and mitigating urban noise pollution. One of its goals is to map the spatiotemporal distribution of noise in real time and over the years in large cities such as New York. In order to reduce noise pollution, citizen participation is crucial, but some residents are unlikely to file a complaint with the city officials. Therefore, the goal of the DCASE 2019 task 5 urban Sound tagging (UST) is to predict whether there are 23 sources of noise pollution in the 10 second scene recorded by the acoustic sensor network. Qiuqiang Kong mixed the datasets of DCASE 2018 Task 1 and Task 2 and proposed a time-frequency segmentation network for sound event detection [1]. In the case of low signal to noise ratio, the better result is obtained. Urban audio contains noise, so we use time-frequency segmentation attention networks for audio tagging.

2. METHOD

2.1. Time-frequency segmentation

This section improves a time-frequency mask algorithm that recognizes and enhances sound events in an audio scene only by implicitly learning in the training of the audio clip. The time-frequency segmentation mask is modeled by convolutional neural network (C-NN), which captures local features of the spectrogram. Each layer of convolutional neural networks includes linear convolution, group normalization (GN) and Rectified linear unit (Relu) activation function. GN is used between linear convolution and group normalization. GN can group channels into groups and calculate the normalized mean and variance in each group. GN calculation have no relation with batch size, so its accuracy is relatively stable in various batches [2]. The last layer of CNN with a Sigmoid nonlinear activation function can output time-frequency segmentation mask, so the output value is between 0 and 1.

2.2. global weighted rank pooling (GWRP)

In order to test whether the segmentation mask can aggregate the segmentation score into the classification score, a new aggregation technique-global weighted rank pooling is used. GWRP calculates the weighted average score of each category to make the most significant part of the weight highest [3]. It can pay more attention to the unit value of the high time-frequency segmentation mask, and pays little attention to unit value of the time-frequency segmentation mask. The GWRP is to put the descending order weights on the values of the time-frequency segmentation mask sorted in descending order. GWRP is defined as:

$$F(m_n) = \frac{1}{Z(r)}\sum_{j=1}^{K} r^{j-1}(m_n)_{i_j} \tag{1}$$

where time-frequency segmentation mask $m = [m_1, ..., m_n]$, N is the number of sound events. The index set $I^C = \{i_1, ...i_K\}$ defines the descending order of the value of the time-frequency segmentation mask m_n, i.e.$(m_n)_{i_1} \geq (m_n)_{i_2} \geq ... \geq (m_n)_{i_K}$. $K = T \times F$ is the number of time-frequency units in the time-frequency segmentation mask. The r is a hyperparameter that can vary according to the frequency of occurrence of the sound event and the constraint range of r is between 0 and 1. $Z(r) = \sum_{j=1}^{K} r^{j-1}$ is a normalization term.

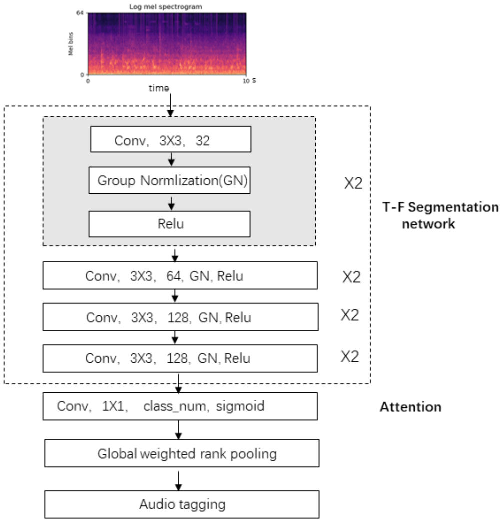

Figure 1: Time-frequency segmentation attention neural network(TFSANN).

2.3. Model

In this section, we propose to the time-frequency segmentation attention neural network for urban sound tagging, which is shown in Fig. 1. The audio clip is first converted into a log mel spectrogram, and then the log mel spectrogram is input into the time-frequency segmentation attention network. Finally the probability value of the audio tagging prediction is obtained.

The time-frequency segmentation attention network consists of four time-frequency segmentation modules, an attention module and a global weighted rank pooling module. Each of time-frequency segmentation module consists of two-layer convolutional neural network. Each layer of convolutional neural network includes linear convolution of filter size 3 × 3, group normalization and Relu activation function. The number of feature maps of convolutional layers are 32, 64, 128 and 128 respectively, and the output of time-frequency segmentation network is feature maps.

The attention module is a 1 × 1 convolution layer with a sigmoid nonlinear activation function, which can convert the feature maps into time-frequency segmentation mask of the sound event. The number of feature maps of the convolutional layer is the number of urban sound categories. When the urban sound is coarse-grained, the number of feature maps is 8. When the urban sound is fine-grained, the number of feature maps is 29. The output of the attention module is a time-frequency segmentation mask.

In order to preserve the resolution of the input time-frequency split mask, the downsampling is not used after the convolutional layer. Finally, the time-frequency segmentation mask is input into the global weighted rank pooling module. The global weighted rank pooling can summarize each time-frequency segmentation mask into a scalar value of the probability of the sound event in the audio clip, and obtain the probability value of the final predicted audio tagging.

The error is calculated by comparing the final output audio label prediction value p_n with the real label value y_n. We use binary cross entropy as a loss function to calculate the error, which can be defined as:

$$E = -\sum_{n=1}^{N}(\boldsymbol{y}_n log \boldsymbol{p}_n) \tag{2}$$

where y_n ,$n = 1, , N$ are limited between 0 and 1, which is the binary representation of the label. $p_n, n = 1, , N$ is the probability of existence of each time-frequency segmentation mask mapped to the n-th event.

3. EXPERIMENTAL RESULTS

3.1. feature extraction

The dataset contains 2351 training samples, 443 validation samples and 274 evaluation samples. The feature extraction method is the Log mel spectrogram [4]. First, the audio clip is sampled at 32KHz, the window size is 1024. Then the short-time Fourier transform is used to obtain the spectrogram of the audio clip. The 64 mel filter bank is applied to the spectrogram to filter the spectrogram. The spectrogram is multiplied by the mel filter bank, and logarithm operation is performed to obtain Log mel spectrogram. Finally each 10s audio clip generates a 640 × 64 feature vector.

3.2. result

Tables 1 and table 2 list the validate average precision of the TF-SANN model for coarse-grained and fine-grained taxonomy on the development dataset. Table 3 gives the validation results of the audio tagging. From the results in the Table 3, the Micro AUPRC, the Micro F1 score and Macro AUPRC scores with coarse-grained taxonomy are better than the baseline system. In the fine-grained taxonomy, Micro AUPRC and Macro AUPRC scores are higher than the baseline system and the Micro F1 score score is lower than the baseline system.

Table 1 Coarse-level validate average precision

Coarse label	Validate average precision
engine	0.765
machinery-impact	0.389
non-machinery-impact	0.321
powered-saw	0.788
alert-signal	0.867
music	0.324
human-voice	0.946
dog	0.103
Avg	0.565

Table 2 Fine-level validate average precision

Fine label	Validate average precision
small-sounding-engine	0.119
medium-sounding-engine	0.543
large-sounding-engine	0.679
engine-of-uncertain-size	0.135
rock-drill	0.239
jackhammer	0.168
hoe-ram	0.137
pile-driver	0.007
other-unknown-impact-machinery	0.432
non-machinery-impact	0.289
chainsaw	0.444
small-medium-rotating-saw	0.477
large-rotating-saw	0.375
other-unknown-powered-saw	0.402
car-horn	0.447
car-alarm	0.101
siren	0.915
reverse-beeper	0.771
other-unknown-alert-signal	0.208
stationary-music	0.293
mobile-music	0.060
ice-cream-truck	0.007
music-from-uncertain-source	0.181
person-or-small-group-talking	0.933
person-or-small-group-shouting	0.397
large-crowd	0.374
amplified-speech	0.010
other-unknown-human-voice	0.061
dog-barking-whining	0.183
Avg	0.324

Table 3 The validation result of audio tagging

		Fine grained	
	Micro AUPRC	Micro F1 score	Macro AUPRC
baseline	0.672	0.502	0.427
TFSANN	0.673	0.376	0.465
		Coarse grained	
	Micro AUPRC	Micro F1 score	Macro AUPRC
baseline	0.742	0.507	0.530
TFSANN	0.802	0.538	0.614

4. REFERENCES

[1] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang and Mark D. Plumbley, “Sound event detection and timeCfrequency segmentation from weakly labelled data,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, No.4, pp. 777–787,2019.

[2] Yuxin Wu and Kaiming He, “GroupNormalization,” arXiv preprint arXiv:1803.08494, 2018.

[3] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 695C711.

[4] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang and Mark D. Plumbley, “Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems,” arXiv preprint arXiv:1904.03476, 2019.