Multi-Scale Time-Frequency Attention for Rare Sound Event Detection

Jingyang Zhang¹, Wenhao Ding¹, Jintao Kang²,³, Liang He¹

¹Department of Electronic Engineering, Tsinghua University, China
²Institute of Forensic Science, Ministry of Public Security, China
³Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, China

{jingyang15,dwh14}@mails.tsinghua.edu, kangjintao@cifs.gov.cn, heliang@mail.tsinghua.edu

Abstract

Attention mechanism has been widely applied to various sound-related tasks. In this work, we propose a Multi-Scale Time-Frequency Attention (MTFA) module for sound event detection. By generating an attention heatmap, MTFA enables the model to focus on discriminative components of the spectrogram along both time and frequency axis. Besides, gathering information at multiple scales helps the model adapt better to the characteristics of different categories of target events. The proposed method is demonstrated on task 2 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge. To the best of our knowledge, our method outperforms all previous methods that don’t use model ensemble on development dataset and achieves state-of-the-art on evaluation dataset by reducing the error rate to 0.09 from 0.13. This demonstrates the effectiveness of MTFA on retrieving discriminative representations for sound event detection.

Index Terms: time-frequency attention, multi-scale learning, sound event detection, DCASE2017 Challenge

1. Introduction

Acoustic Event Detection (AED) defines a task which requires identification of whether some certain types of events occur in an audio clip, and if they do exist, their timestamps should be given. Aiming to help intelligent systems better understand and interact with the surrounding environment based upon acoustic information, AED has drawn great attention among researchers due to its importance. When visual cues are not available or not sufficient, AED is beneficial and even crucial for a system to be aware of current scenarios it is facing. For example, an autonomous car should be able to detect sirens to give way for an ambulance approaching from behind [1]. AED has also been studied in cases like household robots [2] and road surveillance systems [3].

To facilitate the research of rare sound event detection, the task 2 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge [4] provides datasets and a well-designed testbench. The task asks participants to locate the onset of three classes of audio events (babycry, glassbreak, gunshot) within 30-second audio clips, which comprise various acoustic scenes as background.

As deep learning techniques have shown rapid growth and great performance, 4 out of the top-5 entries of DCASE2017 Task2 use deep neural networks to perform rare event detection and indeed achieve good results. Among deep learning-based methods, the combination of Convolutional and Recurrent Neural Network, namely CRNN, is the dominant model architecture. Both 1st place [5] and 2nd place [6] of Task2 utilized CRNN to generate frame-level predictions, i.e., the frame-wise probability of containing rare events. However, target rare events last for only about 2 seconds in the 30-second background and each clip contains at most one event. Therefore, when trained on the frame level, the native CRNN model might be overwhelmed by this data imbalance, i.e., it might focus more on background frames and cannot capture the characteristics of target events. Kao et al. [7] applied Region Proposal Network with CRNN as feature extractor to perform event-level detection, but the performance is highly dependent on the pre-training of CRNN on the frame level according to their report, which doesn’t actually circumvent the problem.

To alleviate the above issue and improve the performance, Shen et al. [8] proposed a temporal attention model and a frequentential attention model to focus on important frames as well as important frequency components. While their attention modules operate separately at different stages of the model propagation, inspired by [9], we propose to leverage temporal and frequentational attention mechanisms within a single mapping using a Multi-Scale Time-Frequency Attention (MTFA) module. MTFA is designed to extract powerful attention-aware features by generating an attention heatmap which tells the model where to focus along both time and frequency axis. Besides, as indicated by its name, MTFA gathers information at multiple scales, enabling the model to zoom in or out and better adapt to different characteristics of different target events. To the best of our knowledge, the proposed method outperforms all previous single-model methods that are without model ensemble on development dataset of DCASE2017 Task2 and achieves state-of-the-art on evaluation dataset. The result provides clear evidence that MTFA has a stronger ability to obtain discriminative representations than previous methods.

2. Method Description

We present an overview of the proposed model in Figure 1. The network consists of three modules: Multi-Scale Time-Frequency Attention, Recurrent Neural Network, and fully-connected layers. MTFA extracts multi-scale attention-aware features with a focus on important parts of the spectrogram. RNN captures temporal relations between each frame. Fully-connected layers produce frame-level predictions which indicate whether the target event occurs within each frame.

2.1. Input Acoustic Features

Mel-spectrogram filtered with log-scale filter banks has been widely used as the raw input for various audio-related tasks. Similar to previous works [5, 6, 7, 10, 9], we extract this feature by first applying a 40ms sliding window with a 20ms shift over the audio signal and then calculating 128-dimensional log filter
bank energies for each frame. Finally, we will get a 2D time-frequency representation with a size of $(1501 \times 128)$ for each 30-second sound clip.

### 2.2. Multi-Scale Time-Frequency Attention

Inspired by Residual Attention which is proposed by Wang et al. [9] for the task of image classification, we design our MTFA module as the feature extractor instead of using a vanilla CNN. The module includes a feature branch for feature extraction and a mask branch for the generation of attention weights.

We choose ResNet [11] consisting of two residual blocks to construct the feature branch. For the mask branch, we utilize a bottom-up top-down structure called Hourglass [12] which has shown its effectiveness in human pose estimation and image segmentation. Hourglass aims to fuse information from multiple resolutions. While encoding local information by downsampling feature maps to smaller and smaller scales along both time and frequency axis, Hourglass also gathers global information through skip connection to maintain high-resolution cues (the residual block within each skip connection are not shown in Figure 1 for brevity). In this work we use 3 Max-pooling layers with kernel size and stride of $2 \times 2$ for downsampling, which means we extract features at 4 different scales: $(T \times D)$, $(T/2 \times D/2)$, $(T/4 \times D/4)$, $(T/8 \times D/8)$. $T$ and $D$ correspond to the length of MTFA’s input along time and frequency axis, respectively. Then, low-resolution feature maps are upsampled (here we use Nearest-Neighbor interpolation) back to the original scale. Via element-wise addition, smaller-scale features are merged with bigger-scale representations which come from the skip connection. Finally, the sigmoid function is applied to generate an attention heatmap whose elements range within $[0, 1]$.

Suppose the input of MTFA has a size of $(C \times T \times D)$, where $C$ denotes the number of feature channels within MTFA which is shared by the feature branch and the mask branch. The output $M$ of the mask branch is of the same size as the representation $F$ produced by the feature branch, which is also $(C \times T \times D)$. To get attention-aware features, one can simply perform element-wise multiplication between $M$ and $F$, i.e., $M \times F$. However, since $M$ ranges from 0 to 1, this naive production might degrade the values of features and make it difficult for the following propagation. Thus, following [9], we use the residual connection and set the output $A$ of MTFA module to be

$$A(x) = (1 + M(x)) \times F(x),$$  \hspace{1cm} (1)

where $x$ represents the input of MTFA module.

Note that while MTFA propagates on a multi-channel level so that it can capture rich information, the input acoustic spectrogram has only one channel. Therefore, before MTFA we use a convolutional layer whose kernel size is $1 \times 1$ together with batch normalization [13] and ReLU to align this mismatch by altering only the channels of the feature map, i.e., from $(1 \times T \times D)$ to $(C \times T \times D)$. A same set of operations is used to get back one-channel features from the output of MTFA before further propagation into the following RNN module.

The underlying intuition of MTFA is to produce powerful temporal-frequential attention-aware features, whose important elements are amplified and noisy parts are suppressed. With these features, the model could not only focus more on the timestamp of target events but also pick out frequency components that are more discriminative than other models without attention. Shen et al. [8] also attempted to utilize temporal-frequential attention, but our method is quite different from theirs. They actually designed two attention models which work separately for time and frequency. Each model consists of a fully-connected layer which takes as input each frame or each frequency component to produce attention weights. On the contrary, by utilizing 2D convolution instead of fully-connected layers, the mask branch of our MTFA module takes in a whole spectrogram and generates a heatmap which directly corresponds to the time-frequency representation extracted by the feature branch. Therefore, the attention weights are element-wise in our method rather than frame-wise or frequency-wise like [8]. In other words, MTFA leverages the 2D nature of convolution and time-frequency features, thus enabling itself to take care of temporal and frequential attention simultaneously within a single mapping. Besides, as convolution operates with $C$ channels ($C > 1$), the mask branch has the capacity and potential to generate attention with richer information than the attention model using fully-connected layers. Thus, we believe

---

**Figure 1:** Global structure of the proposed method. Each green block represents a feature map or the raw input. We first use a convolution operation to transform the input spectrogram to a multi-channel feature map, which is later fed into a MTFA module. MTFA consists of a feature branch for feature extraction and a mask branch for attention generation. The output of two branches is combined in a residual manner to produce multi-scale attention-aware features. Another convolution is applied to get one-channel features back before we pass them into a bi-directional GRU to capture temporal dependencies between frames. Finally, we use fully-connected layers to get frame-wise predictions. The multiplication and addition here are both element-wise.
the proposed method is more straightforward and powerful to perform attention techniques, which will also be demonstrated by experimental results in the following section.

Furthermore, we believe MTFA is beneficial also due to the fusion of multi-scale information. As reported by [8], the average duration of events varies among three target classes: 2.25s for babycry, 1.16s for glassbreak, and 1.32s for gunshot. This is quite intuitive as the latter two events mainly consist of short-time, high-frequency impulses, and babycry generally lasts for a longer time. If the model observes these three classes with one universal time resolution, it might miss either the global information for relatively long-lasting babycry or the local information for transient glassbreak and gunshot. Wang et al. [10] proposed to extract multi-resolution features along the time axis to improve the robustness against time variations. In this work, however, we generate multi-scale representations along both time and frequency axis to further help the model adapt to the characteristics of each rare event, as different events also possess different frequential properties.

### 2.3. Recurrent Neural Network

We use a RNN to capture temporal dependencies between frames, which has been proved to work well when being combined with CNN architecture. Here we use a bi-directional Gated Recurrent Unit (GRU) which has \( U \) hidden units and 2 layers, leading the size of the final high-level feature map to be \((T \times 2U)\). Fully-connected layers and sigmoid function are applied to render frame-wise predictions.

### 3. Experiments

#### 3.1. Dataset

We empirically evaluate the proposed method on the aforementioned task 2 of DCASE2017 Challenge [3]. The dataset consists of 30-second sound clips, each of which has one isolated rare events (either babycry, glassbreak, or gunshot) along with recordings of different acoustic scenes taken from TUT Acoustic Scenes 2016 [14] as the background. We use the provided synthesizer to generate mixtures with different event-to-background ratios and random onset time. For the train set of development dataset, we generate 5000 mixtures for each target event with event-to-background ratios of -6, 0, 6dB and event presence probability of 0.99 in order to have more positive samples. For the test set, 500 mixtures for each target event are generated following the same parameters for evaluation dataset. Note that it is allowed by DCASE2017 to treat target event categories independently. Thus, each type of events will have its own corresponding model, which is also the strategy adopted by all previous works.

#### 3.2. Experiment Setup

During the training phase, the model takes as input a chunk of the spectrogram, i.e, several consecutive frames, rather than the whole spectrogram for the purpose of increasing batch size and training speed. In our experiments, the temporal length \( T \) is set to 256 and the shift between each chunk is 128. During the inference phase, we simply pass the spectrogram as a whole into the network. Since the input will be downsampling trice during propagation, we pad the spectrogram with the last frame to satisfy this length constraint. The number of channels within the MTFA module \( C \) is set to 64, and the number of GRU units \( U \) is also 64. We apply dropout with a rate of 0.3 for babycry and glassbreak, and 0.4 for gunshot as we observe more severe overfitting with it.

The task could be regarded as a 0/1 classification since our method makes frame-wise predictions. Therefore, we use binary cross-entropy as the loss function. We update parameters using Adam [15] with 0.001 as the learning rate. The training is terminated when the validation loss has stopped improving for 10 epochs, and all experiments finished training within 30 epochs according to our observation. The model with the lowest validation loss is selected for the test. We implement our models with PyTorch [16] and perform training with a single NVIDIA TITAN V GPU.

When the model is to be evaluated, the output occurrence probabilities are binarized with thresholds of 0.4, 0.2, and 0.4 for babycry, glassbreak, and gunshot, respectively. The binary presence predictions are further processed with a 540ms median filter, which follows the default setting of DCASE2017 baseline.

#### 3.3. Evaluation Metrics

Event-based error rate and F1-score are two metrics used for DCASE2017 Task2. They are calculated using onset-only condition with a collar of 500ms. The evaluation is done automatically using sed_eval toolbox provided by the task organizers. Details can be found in [17].

#### 3.4. Results

**Comparison with previous methods.** We present results (event-based error rate and F1-score) in Table 1. On development dataset, the proposed method achieves equally as low error rate as R-CRNN [7] on gunshot and also gets desirable results.
on babycry and glassbreak. Note that 1D-CRNN [5] adopted an ensemble strategy and fused the output of up to 5 models to get final predictions. However, our method could perform comparatively well without ensemble and achieves the best average result compared with other single-model methods like [7, 10] and especially [8] which also generates temporal-frequential attention but in a different manner. Note that for all methods, the performance on gunshot is the worst because the sound varies with different gun types. We believe this relatively poor result could be alleviated by incorporating more gun types in the training data.

On evaluation dataset, our method outperforms all previous methods on each target event. With the average error rate slightly increases from 0.07 on development set to 0.09 on evaluation set, it is confirmed that MTFA has a better generalization ability. In a nutshell, the above results demonstrate that MTFA is more powerful an option than previous methods [5, 6, 7, 10, 8] to achieve discriminative features.

Ablation study. An ablation study is conducted to identify the contribution of the generated multi-scale time-frequency attention. We combine ResNet [11] with RNN, which is termed ResNetRNN, to construct a model that has every component of the proposed model in Figure 1 except the mask branch. Hyper-parameters of the model architecture, e.g., the number of channels and residual blocks, are all set to the same as the proposed method for fair comparison. The results are given in the first row of Table 1. While ResNetRNN achieves competitive results on development set, its performance degrades a lot on the evaluation set, indicating that it is insufficient to extract powerful features that could generalize well. By contrast, with the addition of the mask branch, the proposed method improves the performance on both development and evaluation dataset, which again proves the effectiveness of MTFA.

3.5. Visualization of MTFA

We visualize the output of the MTFA module in Figure 2 to provide a straightforward and intuitive understanding. A sound clip from evaluation dataset is taken as an example. As shown by the spectrogram (a), the audio clip comprises the sound of a man’s cough at around 5s, the sound of glassbreak at around 15s, and three sounds similar to tapping the keyboard near the end. The frequency increases from the top to the bottom of each feature map/spectrogram.

4. Conclusion

In this work, we propose a Multi-Scale Time-Frequency Attention module for acoustic event detection, attempting to help the model focus more on discriminative components along both time and frequency axis and extract powerful attention-aware features. We evaluate the proposed method on DCASE2017 Task2. It turns out that our method is the best-performing single-model method on the development dataset, and achieves state-of-the-art on the evaluation dataset. We intuitively and empirically demonstrate that MTFA is a straightforward yet effective option to leverage attention mechanism. We hope this work can provide insight for researchers on adopting multi-scale attention-aware features for acoustic-related tasks. In addition to sound event detection, we believe MTFA could also be applied to acoustic scene classification, audio tagging, etc.

5. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61403224 and No. U1836219.
6. References

[1] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, “Automatic acoustic siren detection in traffic noise by part-based models,” in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 493–497.

[2] X. Wu, H. Gong, P. Chen, Z. Zhi, and Y. Xu, “Intelligent household surveillance robot,” in 2008 IEEE International Conference on Robotics and Biomimetics. IEEE, 2009, pp. 1734–1739.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Audio surveillance of roads: A system for detecting anomalous sounds,” IEEE transactions on intelligent transportation systems, vol. 17, no. 1, pp. 279–288, 2016.

[4] 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 the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), November 2017, pp. 85–92.

[5] H. Lim, J. Park, and Y. Han, “Rare sound event detection using 1D convolutional recurrent neural networks,” DCASE2017 Challenge, Tech. Rep., September 2017.

[6] E. Cakir and T. Virtanen, “Convolutional recurrent neural networks for rare sound event detection,” DCASE2017 Challenge, Tech. Rep., September 2017.

[7] C.-C. Kao, W. Wang, M. Sun, and C. Wang, “Rcrnn: Region-based convolutional recurrent neural network for audio event detection,” in Proc. Interspeech 2018, 2018, pp. 1358–1362. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2323

[8] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, “Learning how to listen: A temporal-frequential attention model for sound event detection,” arXiv preprint arXiv:1810.11939, 2018.

[9] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[10] W. Wang, C.-C. Kao, and C. Wang, “A simple model for detection of rare sound events,” in Proc. Interspeech 2018, 2018, pp. 1344–1348. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2338

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[12] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in European Conference on Computer Vision. Springer, 2016, pp. 483–499.

[13] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[14] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” in 2016 24th European Signal Processing Conference (EUSIPCO). IEEE, 2016, pp. 1128–1132.

[15] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[16] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in NIPS-W, 2017.

[17] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” Applied Sciences, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162