SOUND EVENTS DETECTION AND DIRECTION OF ARRIVAL ESTIMATION USING RESIDUAL NET AND RECURRENT NEURAL NETWORKS

Technical Report

Rishabh Ranjan[1], Sathish s/o Jayabalan[1], Thi Ngoc Tho Nguyen[1], Woon-Seng Gan[1]

[1] Nanyang Technological University, Singapore, 679798
{rishabh001, sathishj, nguyenth003, ewsgan}@ntu.edu.sg

ABSTRACT

This paper presents deep learning approach for sound events detection and localization, which is also a part of detection and classification of acoustic scenes and events (DCASE) challenge 2019 Task 3. Deep residual nets originally used for image classification are adapted and combined with recurrent neural networks (RNN) to estimate the onset-offset of sound events, sound events class, and their direction in a reverberant environment. Additionally, data augmentation and post processing techniques are applied to generalize the system performance to unseen data. Using our best model on validation dataset, sound events detection achieves F1-score of 0.89 and error rate of 0.18, whereas sound source localization task achieves angular error of 9 degree and 0.90 frame recall.

Index Terms— Sound events detection, source localization, residual net, recurrent neural networks

1. INTRODUCTION

Sound events localization and detection system allows one to have automated annotation of a scene in spatial dimension and can assist stakeholders to make informed decisions. It is an important tool for various applications like identifying critical events like gunshots, accidents, noisy vehicles, mixed reality audio where spatial scene information enhanced the augmented listening, robots that listens just like humans and tracks the sound source of interest.

In this paper, we propose a residual net (ResNet) combined with recurrent neural networks (RNN) for the joint estimation of respective labels for sound events detection (SED) and direction of arrival (DoA) for sound events in a reverberant scene with one or two active sound sources. The proposed model outperforms the baseline model [1] using convolutional recurrent neural network (CRNN) especially for DoA task estimation. In the next section, we explain the system configurations followed by results in Section 3. Section 4 describes the DCASE task 3 submission details and concluded in Section 5.

2. MODEL CONFIGURATION

For joint estimation task of SED and DoA, a modified version of ResNet architecture combined with RNN is used. The ResNet model is adapted from residual net model originally designed for image recognition and described in [2].

2.1. Development Dataset

The development dataset consists of 4 splits and each split contains 100 audio files of length 60 sec and contains overlapping and non-overlapping sound events. Audio files is synthesized using 11 isolated sound labels taken from [3] and convolved with impulse response (IR) measured from 5 different rooms at 504 unique combinations of azimuth-elevation-distance and finally, mixed with natural ambient noise collected at IR recording locations.

2.2. Feature Extraction

Each of the audio file is sampled at 48kHz and short-time Fourier transform (STFT) is applied with hop size of 20 msec. Next, STFT spectrogram is converted to log mel spectrogram for magnitude, while mel filter banks used to convert phase spectrogram into mel spectrogram. After converting into mel spectrogram features, low and high frequency components are removed and finally, resized to match the input shape of the neural network before training.

2.3. Model Training

For development, 4 cross-fold sets from detection and classification of acoustic scenes and events (DCASE) challenge 2019 task 3 [4] is used as recommended by DCASE organizers. Each cross-fold consists of 2 training split, 1 validation split, and 1 test split as shown in Table 1. For evaluation and DCASE submission, we follow 3-splits for training and one split for validation as shown in Table 2. During training, each processed audio feature file is split into sequence length of 128 frames and resized with fixed batch size of 96. The batch-feature dimension is $Batch\_size \times N_{ch} \times Seq\_length \times N_{mel}$, where $N_{ch} = 8$ corresponding to 4 channels for magnitude and 4 channels for phase concatenated together. N_{mel} is the number of filter banks and is varied between 64 and 128.

Table 1: Cross-fold configuration for model development

Fold	Training sets	Validation sets	Test sets
1	Split 3, 4	Split 2	Split 1
2	Split 4, 1	Split 3	Split 2
3	Split 1, 2	Split 4	Split 3
4	Split 2, 3	Split 1	Split 4

Table 2: Cross-fold configuration for model evaluation

Fold	Training sets	Validation sets
1	Split 2, 3, 4	Split 1
2	Split 3, 4, 1	Split 2
3	Split 1, 2, 4	Split 3
4	Split 1, 2, 3	Split 4

For SED, binary cross-entropy loss function with sigmoid activation function in output layer is used for multi-label classification with 11 classes. For DoA, weighted binary cross-entropy loss function so as to strongly penalize the false negatives because most of the ground truth labels are zero. DoA predictions are summarized as probabilities for 324 classes corresponding to 36 azimuths and 9 elevations. For SED and DoA, adam optimizer is used with learning rate of 0.0005. Best model is saved using the evaluation metrics provided by DCASE task 3 organizers.

2.4. Data Augmentation

To improve model generalization capability on unseen test data, data augmentation using frame shifting is applied to each of the processed audio file. Each audio file with 3000 frames is shifted by 32, 64 and 96 frames in temporal dimension before splitting into sequence of 128 frames. Therefore, total data after augmentation is 4 times larger than the original dataset size and is selected randomly for training in each epoch.

2.5. Evaluation Metrics

Model performance is evaluated using 4 metrics, 2 each for SED and DoA. SED is evaluated using error rate (ER) and F-score. ER is the total error based on total number of insertions (I), deletions (D) and substitutions (S) [5]:

$$ER = \frac{S + D + I}{N}, \quad (1)$$

where N is total number of frames. F-score is calculated as harmonic mean of precision (P) and recall (R) [5]:

$$F - Score = \frac{2P \cdot R}{P + R} \quad (2)$$

DoA is evaluated using average angular error and frame recall. DoA error is defined as average angular error in degrees between estimated and ground truth directions and computed using Hungarian algorithm [6] to account for the assignment problem of matching the individual estimated direction with respective reference direction. DoA frame recall (FR) is defined as percentage of frames where number of estimated and reference directions are equal.

3. RESULTS

Table 3 shows the model performance of proposed ResNet RNN compared to baseline CRNN model [1] for 3-split model training configurations as described in Table 2. Clearly, the proposed model outperforms the baseline model in terms of all the 4 metrics and all folds. There is significant improvement in terms of DoA error from 28° for baseline to 9° for the proposed method.

Table 3: Proposed ResNet RNN model Vs Baseline CRNN model performance fold-wise

Fold	Model	ER	F-Score	DoA Error (°)	FR (%)
1	Proposed	**0.1644**	**89.85**	**9.53**	**91.84**
	Baseline	0.3055	82.96	30.13	85.42
2	Proposed	**0.1911**	**89.12**	**9.29**	**90.11**
	Baseline	0.3273	82.17	31.32	82.44
3	Proposed	**0.1617**	**90.69**	**8.50**	**91.36**
	Baseline	0.2676	84.93	31.75	86.12
4	Proposed	**0.2146**	**86.79**	**8.51**	**90.26**
	Baseline	0.2937	82.64	31.05	87.23
overall	Proposed	**0.1829**	**89.10**	**8.96**	**90.89**
	Baseline	0.2986	83.16	31.06	85.30

Table 4: Proposed ResNet RNN model Vs Baseline CRNN model performance for Ov1 and Ov2

Fold	Model	ER	F-Score	DoA Error (°)	FR (%)
Ov1	Proposed	**0.1589**	**91.23**	**3.80**	**95.60**
	Baseline	0.2834	85.32	26.41	93.23
Ov2	Proposed	**0.1953**	**87.95**	**11.70**	**86.18**
	Baseline	0.3064	82.00	33.70	77.38

Table 4 shows the proposed model performance for single source (Ov1) and two overlapping sources (Ov2). Clearly, model performs much better for single source scenario as compares to two source, especially for DoA error being 3.8° for Ov1 and 11.7° for Ov2 cases. In the next section, we outline the models' description which was submitted to DCASE challenge Task 3.

4. DCASE SUBMISSION

Based on the 2-split and 3-split model results, 4 best models were selected for submission. Best models were decided based on combined score of all the 4 evaluation metrics.

4.1. Submission 1

Model corresponding to *fold 1* in Table 3:

- SED: Model trained using 64 mel bands; magnitude only
- DoA: Model trained using 128 mel bands; both magnitude and phase

4.2. Submission 2

Model corresponding to *fold 3* in Table 3:

- SED: Model trained using 128 mel bands; magnitude only
- DoA: Model trained using 128 mel bands; both magnitude and phase

4.3. Submission 3

Ensemble model of all the 4 models listed in Table 3. Model performance should be closer to overall metric in Table 3.

4.4. Submission 4

All the 4 development splits were used for training and its performance should be as good as our best model in Table 3.

5. CONCLUSION

In this report, ResNet model combined RNN architecture is used for sound events classification and localization task. With data augmentation and post-processing techniques, the proposed model is significantly improved overall, and especially for the source localization with DoA error improvement of more than 20° for both single and two overlapping sources.

6. ACKNOWLEDGMENT

This research was conducted in collaboration with Singapore Telecommunications Limited and supported by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

The authors are also thankful to Maggie Leong at Amazon Web Services (AWS) Singapore to generously provide the resources for model development on the cloud.

7. REFERENCES

[1] Adavanne, Sharath, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. "*Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks.*" IEEE Journal of Selected Topics in Signal Processing (2018).

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "*Deep residual learning for image recognition.*" In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[3] TUT audio dataset: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio#audio-dataset.

[4] DCASE Challenge Task 3: http://dcase.community/challenge2019/task-sound-event-localization-and-detection

[5] Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. "*Metrics for polyphonic sound event detection.*" Applied Sciences 6, no. 6 (2016): 162.

[6] H. W. Kuhn, “The hungarian method for the assignment problem,” *in Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397