WEAKLY LABELED SOUND EVENT DETECTION WITH RESDUAL CRNN USING SEMI-SUPERVISED METHOD

Technical Report

Jie Yan

University of Science and Technology of China
National Engineering Laboratory for
Speech and Language Information Processing
Hefei, China
yanjie17@mail.ustc.edu.cn

Yan Song

University of Science and Technology of China
National Engineering Laboratory for
Speech and Language Information Processing
Hefei, China
songy@ustc.edu.cn

ABSTRACT

In this report, we present our system for the task 4 of DCASE 2019 challenge(Sound event detection in domestic environments). The goal of the task is to evaluate systems with real data either weakly labeled or unlabeled and simulated data that is strongly labeled. To perform this task, we propose resdual CRNN as our system. We also use mean-teacher model based on confidence thresholding and smooth embedding method. In addition, we also apply specaugment for the labeled data shortage problem. Finaly, we achieve better performance than DCASE2019 baseline system.

Index Terms— Sound event detection, resdual CRNN, Smooth embedding, Confidence thresholding

1. INTRODUCTION

Sounds contain a large amount of information about our everyday environment and physical events that take place in it. These information can help humans understand the surroundings even without visual information. Developing systems to automatically extract this information has huge potential in several applications including multimedia indexing [1], intelligent monitoring system in living environment [2], health care [3], etc. DCASE challenge has been organized to make a significant amount of effort to promote developing reliable method for recognition of sound scenes and individual sound sources in realistic soundscapes. Challenge in this year comprises five tasks: acoustic scene classification, audio tagging with noisy labels and minimal supervision, sound event localization and detection, sound event detection in domestic environments, and urban sound tagging. This report describes a solution to task 4 of the DCASE 2019 challenge, which is the follow-up to DCASE 2018 task 4. It is aiming to investigate is whether we really need real but partially and weakly annotated data or is using synthetic data sufficient? or do we need both? We propose several improvement on data augmentation, model architecture and training method to achieve better performance.

2. PROPOSED METHODS

We present specaugment for data augmentation, resdual CRNN model architecture and improvements on mean-teacher training method.

2.1. Data augmentation

As the amount of weakly labeled data used in task 4 is small, we apply specaugment method [4] to generate additional training data for the task. The augmentation improve generalization ability of the model for various unseen data.

To be specific, we construct augmentation policy on the log Melscale spectrogram. The policy is frequency masking applied to the spectrogram so that f mel frequency bands $[f_0, f_0 + f)$ are masked, where f is first selected from a uniform distribution from 0 to the frequency mask parameter F, and f_0 is selected from $[0, v - f)$. v is the length of frequency axis. This policy makes model robust to the partial loss of frequency information.

2.2. Proposed architecture

Our system is based on the baseline system and the best submission of DCASE2018 task4 [5], shown in Fig.1. We inroduce shake-shake mechanism to context gating module for regularization and propose resdual CRNN to make use of both convolutional features and recurrent fetures.

2.2.1. Shake-shake mechanism

Due to the weakly labeled data shortage, we may run into overfit promblem easily. We propose shake-shake mechanism [6] to alleviate the problem. Based on the typical context gating module [7], we propose another branch added to it. The module follows this equation;

$$Y = \sigma(\alpha\mathcal{F}(X, \mathcal{W}_1) + (1 - \alpha)\mathcal{F}(X, \mathcal{W}_2)) \odot X \qquad (1)$$

where X is the input feature, σ is sigmoid activation and $\odot$ is element-wise multiplication. $\mathcal{W}_1$ and $\mathcal{W}_2$ are sets of weights associated with the 2 branches. α is a random variable following a uniform distribution between 0 and 1 during training and sets to the value of 0.5 during testing. The random numbers perform update operation before ezch forward and backward pass. In this way, we apply stochastic disturbance to our system.

2.2.2. Resdual CRNN

CRNN framework [8] has been shown to efficient for the sound event dectection. Based on this, we present resdual CRNN [9] to

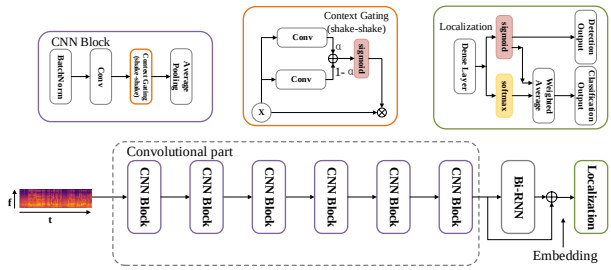

Figure 1: The overall structure of the proposed model.

establish relatonship between local features and contextual features. Let x_t denote the output of CNN part and z_t denote the output of RNN part. The resdual features at time step t is;

$$a_t = \mathrm{ReLU}(x_t \mathcal{W}_x + b_x) \oplus z_t \tag{2}$$

where $\mathcal{W}_x$ and b_x are the weight of a fully-connected layer and $\oplus$ indicates vector concatenation.

2.3. Semi-supervised training methods

We use mean-teacher method [10] to explore possible usage of the large amount of unlabeled data. In this method, the student model is trained on weakly labeled data and weights of the teacher model are a moving average of the student model. A cost for consistency between teacher and student model is applied on both labeled and unlabeled data.

As for consistency loss, we use confidence thresholding method [11] to improve it. For the sample s_i, if the predicted probability of teacher model is below the confience threshold, the consistency loss for the sample s_i is masked to 0. This means if the teacher model doesn't learn well for the sample, it should not teach the student. This help student learn correct labels from the steacher.

We also apply smooth embedding [12] method for our system. Let h denote the mapping from the input space to the embedding feature space which is the output feature map of the resdual CRNN. We define the loss on the embedding;

$$l_S = \begin{cases} \|h(x_i) - h(x_j)\|^2 & \text{if } y_i = y_j \\ \max(0, m - \|h(x_i) - h(x_j)\|)^2 & \text{if } y_i \neq y_j \end{cases} \tag{3}$$

where $m > 0$ is a margin and $\|\cdot\|$ is Euclidean distance. The loss urge smaples with the same label to have consistent embeddings and push samples with different labels apart from each other. It helps obtain more expressive and discriminative representation in a smooth and coherent feature space.

3. EXPERIMENT AND RESULTS

In this section, we show the system setup and report thr performance of our model.

3.1. dataset

The dataset for training comprises 3 parts, including weakly labeled dataset, unlabeled in domain dataset and synthetic strongly labeled dataset. The amout of weakly labeled data is rare. The dataset is composed of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. There are 10 kinds of events including Speech, Dog, Cat, Alarm/bell/ringing, Dishes, Frying, Blender, Running water, Vacuum cleaner and Electric shaver/toothbrush. In addition, the events in audio clip may partly overlap.

Table 1: Experimental results for the baseline and proposed methods.

Model	F1	Precision	Recall
DCASE 2019 Baseline	23.7	-	-
ResCRNN-MT	39.4	39.0	41.6
ResCRNN-MT-fusion	42.1	42.5	42.2

3.2. Experimental setup

Log-Mel spectrogram is extracted from audio clips by 64-bin, 2048-window and 511-hop. The spectrogram used as the input of system has size of 864×64. The model shown in Fig. 1 consists of six CNN blocks in the convolutional part, followed by a bi-RNN. The convolution layers in CNN blocks have 128 channels with 3×3 kernel size. Adam optimiezer is used to optimize the network with a learning rate of 0.001.

3.3. Results

In this section, we compare our system based on resdual CRNN using mean-teacher method with the baseline system. Results presented in Table 1 show that our proposed method performs better than baseline. Moreover, we fuse our model with the model using the audio tagging result as the sound event detection result as presented in [13]. This fusion achieves even better result.

4. CONCLUSION

This report describes some methods for the sound event detection in task 4. The specaugment is used to enlarge data and deal with labeled data shortage problem. More expressive features are obtained by resdual CRNN and mean-teacher method with confidence thresholding and smooth embedding. Finally, our model achieves better resukt than the baseline model. In the future, more efforts will be made to the usage of synthetic strongly labeled dataset for improvement.

5. REFERENCES

[1] D. Zhang and D. Ellis, "Detecting sound events in basketball video archive," *Dept. Electronic Eng., Columbia Univ., New York*, 2001.

[2] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[3] J. Schroeder, S. Wabnik, P. W. Van Hengel, and S. Goetze, "Detection and classification of acoustic events for in-home care," in *Ambient assisted living*. Springer, 2011, pp. 181–195.

[4] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[5] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," *Detection and Classification of Acoustic Scenes and Events*, 2018.

[6] X. Gastaldi, "Shake-shake regularization of 3-branch residual networks," 2017.

[7] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[8] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[9] H. Phan, O. Y. Chén, P. Koch, L. Pham, I. McLoughlin, A. Mertins, and M. De Vos, "Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 51–55.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[11] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," *arXiv preprint arXiv:1706.05208*, 2017.

[12] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8896–8905.

[13] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Dcase 2018 challenge surrey cross-task convolutional neural network baseline," *Parameters*, vol. 4, pp. 4–691, 2018.