MEAN TEACHER WITH SOUND SOURCE SEPARATION AND DATA AUGMENTATION FOR DCASE 2020 TASK 4

Technical Report

Chuming Liang, Haorong Ying, Yueyi Chen, Zhao Wang

Xiaomi AI Lab, Wuhan, CHINA

ABSTRACT

In this paper, we present our system for the DCASE 2020 challenge Task4(Sound event detection and separation in domestic environments). The target of this task is to provide time boundaries of multiple events in an audio recording using a system trained with unlabeled, weakly-labeled and synthetic data. Also, sound source separation is encouraged to use in the system. We propose a mean-teacher model with convolutional and recurrent neural network(CRNN) structure and adopt data augmentation and sound source separation technique to improve the performance of sound event detection.

Index Terms— DCASE2020, sound event detection, mean teacher, sound source separation

1. INTRODUCTION

In this paper, we propose a sound event detection model based on the provided baseline. The baseline uses a mean-teacher model[1] to deal with unlabeled and weakly labeled data. The main architecture of baseline is a convolutional and recurrent neural network adopted from [2]. From the baseline, we use context gating[3] as the activation function in our model and adopt different data augmentation techniques. Further more, we modify the sound source baseline system to achieve better performance.

2. DATASET

The training dataset of DCASE2020 is consisted of three parts: weakly-labeled data, unlabeled data and synthetic data with strong annotations. Weakly-labeled data contains 1578 recordings, unlabeled in domain data contains 14412 recordings and synthetic data contains 2584 recordings. Besides, foreground event sounds and background sound are also provided so that participants can generate more training data. In our system, we only use the provided synthetic data. Weakly-labeled and unlabeled recordings are sampled at 44100Hz while the synthetic data has a sampling frequency of 16000Hz. There are 10 event labels in the dataset: Speech, Dog, Cat, Alarm, Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver.

2.1. Audio preprocessing

Feature extraction settings are the same as baseline system. First, all audio recordings are resampled to 16000Hz. Then we extract the log-mel spectrogram from the audio clips. The number of mel bins is 128. The window length of each frame is 2048(128ms) and the hop length is 255(16ms). We have performed experiments to show that smaller hop length can achieve better SED performance. Finally, we perform mean and variance normalization where the mean and standart deviation value are computed on the entire training set.

3. PROPOSED SOLUTION

Our proposed solution is based on the baseline system, which is a mean-teacher model. The baseline uses two CRNN networks where the second model(teacher model) is the exponential moving average of the first model(student model). The mean-teacher model has achieved the state-of-the-art results of SED since it exploited a large amount of unlabeled data effectiviely. Another semi-supervised method in SED is guided learning[4], which however achieved worse results than the baseline system in experiments. To improve the performance of the baseline system, we use different data augmentation techniques, modify the network structure, apply a specific median filter for each class and integrate sound source separation in the training and evaluation phase.

3.1. Data augmentation

3.1.1. Time stretching and pitch shift

To further enhance the generalization ability of our model, we also propose to introduce data augmentation by shifting the normalized mel-spectrogram along time and frequency axes. The parameters of two functions are set to be random. For time-shifting, the spectrogram is wrapped along the time axis, e.g. for a positive time-shift, the last frames of the spectrogram become the first frames of the shifted spectrogram. Pitch shift is an adjustment of pitch, and the pitch depends on the frequency. The higher the frequency, the higher the pitch. Therefore, pitch shift can be regarded as a scale transformation of frequency.

As mentioned above, the time streching procedure is defined in the next equation:

$$t_{final} = \frac{t_{orig}}{rate} \tag{1}$$

where rate represents the strech factor. If $rate$>1, the signal is sped up, otherwise, the signal is slowed down.

In pitch shifting, we sampled the audio at a sampling rate of 16,000. The signal s(n) can be expressed as the sum of its periodic and stochastic parts:

$$s(n) = \sum_{k=1}^{K} MAG_k(n) cos\phi_k(n) + r(n) \tag{2}$$

where $MAG_k(n)$ is the magnitude of the kth sinusoidal component, K is the number of the components, $\phi_k(n)$ is the instantaneous phase of the kth component, and r(n) is the stochastic part of

the signal. Instantaneous phase $\phi_k(n)$ and instantaneous frequency $f_k(n)$are related as follows:

$$\phi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \phi_k(0) \tag{3}$$

where F_s is the sampling frequency and $\phi_k(0)$ is the initial phase of the kth component.

The musical range is divided in many octaves. Each octave is made of twelve semitones, also referred to as half steps.When using pitch shifting, the semitones will go up or down, as shown in the next equation.s is the number of semitones for shifting.

$$p_{final} = p_{orig} + s \tag{4}$$

From a frequency point of view, the change of frequency can be calculated as

$$f_{final} = 2^{(s/12)} \times f_{orig} \tag{5}$$

3.1.2. Reverb

We also tried to add reverberation to the audio which generally refers the creation of artificial spatial effects.Reverberation can be simulated by using the sum of geometric sequence

$$y(n) = x(n) + ax(n-D) + a^2x(n-2D) + ... \tag{6}$$

where $x(n)$is the original acoustical signal, $y(n)$ is the reverberant signal, $a$is the attenuation coefficient, D refers the delay time. The transfer function can be described as

$$H(z) = 1 + az^{-D} + a^2z^{-2D} + ... \tag{7}$$

The transfer function can be converted into:

$$H(z) = \frac{1}{1 - az^{-D}} \tag{8}$$

We reverb the audio using the library named pysndfx. In this part, the reverberance parameter is set to a random value with a Normal distribution with a 0.3 mean and standard deviation of 0.05.

3.2. Network structure

The network structure is basically the same as the baseline model. The modifications we make are as follows:

1. The RNN layer is a bidirectional 2-layer GRU where each layer contains 256 cells.
2. The number of layer of CNN remains 7, but the setting of filter numbers become: [16,32,64,128,128,256,256]. The kernel size and pooling size are the same as the baseline.
3. The activation function in our system is context gating instead of GLU. Context gating has shown great performance in previous DCASE task[5] with less parameters compared with GLU. The formula of context gating is:

$$Y = \sigma(\omega * X + \beta) \odot X \tag{9}$$

   where X is the input feature vector, ω, β are trainable parameters and $\odot$ is the elementwise multipication and σ is the sigmoid function.
4. For training, we train for 300 epoches with a learning rate of 0.001. The optimizer is Adam and the best model on validation set is selected as the submitted model.

3.3. Sound source seperation

In sound-separation part , we've done a lot of efforts to expand the scale of data. First we change the parameter of scaper to generate a new augmented dataset. Second, we apply time stretch and pitch shift function to the feature map. It should be noted that the parameters of the two functions are random. For more information about functions, see the description below. We linearly scale the time axis of the spectrograms by a factor stretch and keep the central part.Stretch is drawn randomly from a uniform distribution between 0.6 and 1.4 for each sample. Note that this is an approximation compared to an actual modification of the speed of the audio. Time stretching : We linearly scale the time axis of the spectrograms by a factor time stretch which is drawn randomly from a uniform distribution between 0 and 1 for each sample.

Meanwhile , we apply the repeat function to generate infinite training data on paper.

Sound source separation is integrated in the evaluation phase of proposed system as the baseline.

4. RESULTS

In DCASE2020 task4, the event-based F1-score(macro average) is used to evaluate the performance of the system. Table 1 shows the results of our proposed system on validation set for each class.

Table 1: F1-score for proposed system on validation set

Class	F1-score(%)
Alarm/bell/ringing	18.4
Blender	38.8
Cat	42.0
Dishes	29.7
Dog	27.2
Electirc shaver/toothbrush	28.4
Frying	38.6
Runing water	35.3
Speech	53.7
Vacuum cleaner	43.6
Global	35.57

5. CONCLUSION

In this paper, we proposed a sound event detection system based on a mean-teacher model and inspired by the provided baseline. The propsed system achieves better performance than the baseline system thanks to data augmentation techniques, improvement of the network structure and integration of sound source separation. Our proposed system achieved 35.57% of F1-score on the validation set

6. REFERENCES

[1] 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.

[2] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," Orange Labs Lannion, France, Tech. Rep., June 2019.

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

[4] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 626–630.

[5] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.