Weakly and semi-supervised learning for sound event detection using image pretrained convolutional recurrent neural network, weighted pooling and mean teacher method

Xichang Cai\textsuperscript{1*}, Dongchi Yu\textsuperscript{1}, Duxin Liu\textsuperscript{1}, Menglong Wu\textsuperscript{1}
\textsuperscript{1}Department of Electronic Information Engineering, North China University of Technology, Beijing, China
*Corresponding author’s e-mail: caixc20_ncut@126.com

Abstract. In this paper, we propose a sound event detection (SED) method which uses deep neural network trained on weak labeled and unlabeled data. The proposed method utilizes a convolutional recurrent neural network (CRNN) to extract high level features of audio clips. Inspired by the impressive performance of transfer learning in the field of image recognition, the convolutional neural network (CNN) in the proposed CRNN is an image-pretrained model. Although there is a significant difference between audio and image, the image-pretrained CNN still has competitive performance in SED and can effectively reduce the amount of training data needed. To learn from weak labeled data, the proposed method utilizes a weighted pooling strategy which enables the network to focus on the frames containing events in an audio clip. For unlabeled data, the proposed method utilizes a mean teacher semi-supervised learning method and data augmentation technique. To demonstrate the performance of the proposed method, we conduct the experimental evaluation using the DCASE2021 Task4 dataset. The experimental results demonstrate that the proposed method outperforms the DCASE2021 Task4 baseline method.

1. Introduction
In the world we live in, a large number of sound events occur every moment, such as baby crying, car horns and gun shot. The information provided by these sound events can help people understand their surroundings. Sound event detection (SED) is a technique to find out which sound events occurred in an audio recording and locate the onset and offset time of them. SED has been applied in many fields such as environment monitoring\cite{1}\cite{2} and urban sound analysis\cite{3}.

Recently, most SED approaches are based on deep neural networks.\cite{4} introduces a convolutional neural network (CNN) with a large input field to extract the long-time frequency audio feature and enables to train SED end-to-end.\cite{5} presents a multi label bi-directional recurrent neural network (RNN) to model the temporal evolution of sound events.\cite{6}\cite{7} combine the advantages of CNN and RNN in a convolutional recurrent neural network (CRNN). CNN is used to extract the local frequency and time invariance of the spectrum, and RNN is used to model the long term temporal context of audio. However, the above mentioned approaches are using strong label training data, which contains the types and the timestamps of sound events. The annotations of strong label training data are both labor- and cost-intensive for human annotators to produce, which limits the practical scalability of SED methods. In contrast, weakly labeled data which indicates the presence or absence of sound events but no timestamp information and unlabeled data which has no label information are much easier to collect. Therefore,
the research of sound event detection using weak labeled data and unlabeled data has become a hot topic in this field.

Weakly supervised learning[8] is a technique to train deep neural networks using weakly labeled data. One of the common approaches for weakly-supervised SED is multiple instance learning (MIL)[9]. In the MIL, an audio clip is divided into many frames and sent to a classifier to get frame-level prediction (i.e. strong label). Then, clip-level prediction (i.e. weak label) is produced by aggregating all of the frame-level predictions to calculate loss function with weak label. Because sound events may only occur in certain frames, the difficulty in the process of clip level prediction generation is how to make the network pay attention to these certain frames. Similarly, semi supervised learning is a technique to train deep neural networks using unlabeled data. Recently, the key method of semi supervised learning is consistent regularization, whose main target is to obtain perturbation/augmentation invariant output distribution. Therefore, the consistency regularization method highly relies on the progress of data augmentation, such as [10][11]. Since most of the current data augmentation methods are for images, how to use semi supervised learning method and data augmentation technology in audio field is a problem to be studied.

In addition, most of the previous SED methods train neural networks from scratch due to the lack of large-scale audio data sets. According to the research experience in the field of image recognition, transfer learning and pretrained model can significantly improve the network performance on small datasets. [12][13][14] show that the knowledge learned by the model in the image pretraining is also helpful to deal with the problems in the audio field.

In this paper, we propose a weakly and semi-supervised SED method. Our proposed method utilizes an image pretrained CRNN as backbone and introduces a weighted pooling strategy and a mean teacher semi-supervised technique to learn from weakly labeled and unlabeled data. To demonstrate the performance of the proposed method, we conduct the experimental evaluation using the DCASE2021 Task4 dataset. The experimental results demonstrate that the proposed method outperforms the DCASE2021 Task4 baseline method.

2. Materials and Methods

2.1. Dataset

We train and test the proposed model on DCASE2021 Task4 dataset[15]. Detection and Classification of Acoustic Scenes and Events (DCASE) is a challenge that evaluates the development of computational scene and event analysis methods by comparing different approaches using a common publicly available dataset. The DCASE2021 task4 dataset consists of four subsets belonging to 10 classes, including training sets (synthetic strongly labeled: 10,000 clips, weakly labeled: 1,578 clips, unlabeled: 14,412 clips) and validation set (1,168 clips). Most of the clips last 10 seconds. The sampling rate of clips is not uniform, most of them are 16K or 44.1KHz. Multiple audio events may occur at the same time.

2.2. Data pre-processing

Different types of sound events usually have large frequency differences, so the frequency of audio plays an important role in SED. The original audio waveform is one-dimensional signal, which does not contain frequency information explicitly. Therefore, we transform the original audio waveforms into log mel-spectrograms as the input. The log mel-spectrogram is a two-dimensional signal, which has two dimensions of time and frequency. The log mel-spectrogram contains the corresponding relationship between frequency and time.

The process of generating the log mel-spectrogram is as follows. First, the input audio clips are resampled at 16k Hz. A 2048-point hamming window with the hop size of 256 is then adopted to divide the raw audio clips into frames. 2048-point FFT and 128 log-Mel filter banks are used to extract log-Mel feature on each frame. Finally, the 10s raw audio clips are converted to the log mel-spectrogram features with the shape of 626 by 128. In addition, we copy the one-channel log mel-spectrogram into three-channel to adapt to the pretrained CNN models, which usually use three-channel image as input.
2.3. Network architecture

As shown in Figure 1, our proposed network consists of three parts: a CNN feature encoder, a RNN feature encoder, a frame level classifier and a weighted pooling layer.

The input log mel-spectrogram can be formulated as \( X = (x_i | i = 1, ..., T) \), where \( x_i \) is the 128-dimensional feature vector for \( i \)-th frame, and \( T \) is the total numbers of frames. We treat the log mel-spectrogram as an image, and use an image-pretrained CNN to extract local time and frequency invariance of sound events. Due to the limitation of hardware performance, we choose Efficientnet-B0[16] as our CNN feature encoder. Efficientnet-b0 is a pretrained model on Imagenet[17], which achieves a high performance level with greatly reducing the number of parameters. Because the original efficientnet-b0 is designed to process image classification task, the final output feature map of it is 32 times smaller than the input, which will greatly reduce the temporal resolution of log mel-spectrogram in SED task. Therefore, we abandon some deep network layers in efficientnet-b0 to maintain temporal resolution. The CNN feature encoder architecture we used is shown in Table 1. We obtain embedded feature \( E_{CNN} = (e_n^{CNN} | t = 1, ..., T/8) \) by sending the input log mel-spectrogram \( X \) to CNN, where \( e_n^{CNN} \) is the 16-dimensional feature vector for \( n \)-th frame. Note that the input log mel-spectrogram \( X \) and the embedded feature \( E_{CNN} \) have multiple channels, 3 and 40, respectively. The embedded feature \( E_{CNN} \) already contains local time and frequency invariance. However, many types of sound events last for a long time, and long-term context information can help the model detect these sound events better. RNN can extract the context information of time sequences. We use a bi-directional gated recurrent unit (bi-GRU)[18] as RNN feature encoder. The input size and hidden size of bi-GRU are 640 and 128, respectively. We concatenate the channel dimension of \( e_n^{CNN} \) along the frequency dimension. Then we obtain temporal embedding feature \( E_{RNN} = (e_t^{RNN} | t = 1, ..., T/8) \) by sending \( e_n^{CNN} \) into RNN feature encoder, where \( e_t^{RNN} \) is the 256-dimensional feature vector for \( n \)-th frame. Finally, the classifier with a fully connected layer and Sigmoid activation function is applied to the temporal embedding feature \( E_{RNN} \), and then the classification prediction of each frame (i.e. strong label prediction) \( p_{strong} = (p_t | t = 1, ..., T/8) \), where \( p_t \) is 10-dimensional vector representing the probability of each type of sound event. The classification prediction of the whole audio (i.e. weak label prediction) is obtained by weighted pooling, which will be explained in detail in the next section.

| Block                  | Output size | Output channels | Layers |
|------------------------|-------------|-----------------|--------|
| Conv,k3x3,s2x2         | input/2     | 32              | 1      |
| MBConv1,k3x3,s1x1      | input/2     | 16              | 1      |
| MBConv6,k3x3,s2x2      | input/4     | 24              | 2      |
| MBConv6,k5x5,s2x2      | input/8     | 40              | 2      |

2.4. Weighted pooling

In order to train SED system with weak label data, the frame level prediction output by the system needs to be aggregated into clip level prediction to calculate the loss function. Frames with target sound events are expected to be more important than other frames. We introduce a temporal attention mechanism and a weighted pooling strategy to address this problem.

The temporal attention mechanism is formulated as follows:

\[
a_{ij} = \text{softmax}(W_2^T \tanh(W_1 F_{ij}^T + b))
\]  

(1)
\[
P = \sum_i \sum_j a_{ij} p_{ij}
\]

where \(W_1\) and \(W_2\) are trainable weight parameter matrices of this two-layer attention network, and \(b\) is the bias parameter matrix. \(F^l\) is the frame level feature vector, \(a_{ij}\) is the weight corresponding to frame level prediction \(p_{ij}\) and \(P\) is the final clip level prediction.

The weighted pooling strategy is formulated as follows:

\[
P = \frac{\sum_i \sum_j p_{ij} \times p_{ij}}{\sum_i \sum_j p_{ij}}
\]

in which there are no trainable parameters. Different temporal attention mechanism, this strategy is more interpretable. The partial derivative of \(P\) to \(p_{ij}\) is as follows:

\[
\frac{\partial P}{\partial p_{ij}} = \frac{2p_{ij} - P}{\sum_i \sum_j p_{ij}}
\]

When \(2p_{ij} > P\), \(p_{ij}\) will be increased. On the contrary, \(p_{ij}\) will be decreased. We will compare the performance of the two methods in the experimental part.

2.5. Semi-supervised learning

To learn from unlabeled training data, we implement a semi-supervised learning method called mean teacher[19]. The mean teacher method consists of two network model called student model and teacher model respectively. Student and teacher model share the same structure of our proposed multi-scale CRNN. The weights of the student model are updated with gradient back propagation, and the weights of the student model are updated as an exponential moving average (EMA) of the student weights. The same data is input into two models, and the network parameters are optimized according to the consistency regularity of the two network outputs.

2.6. Data augmentation

We employ mixup[20] for data augmentation. Mixup can improve the performance of deep neural network in many machine learning tasks by smoothing the distribution of samples in the feature space. This method creates a new data by interpolate between two raw data, while the labels are interpolated in the same way. This process is expressed as

\[
\bar{x} = \lambda x_i + (1 - \lambda)x_j
\]

\[
\bar{y} = \lambda y_i + (1 - \lambda)y_j
\]

where \(x_i, x_j\) are different data points, and \(y_i, y_j\) are corresponding labels.

3. Results & Discussion

3.1. Training

The Adam optimizer and learning rate of 0.001 are used for training. The batch size and number of epochs are 48 and 200 respectively. We employ the exponential warmup strategy to gradually increase the learning rate from very small to 0.001 in the first 50 epochs, the learning rate remained unchanged during subsequent training. All the models were trained using a single Nvidia GTX 1080 GPU.

3.2. Evaluation metrics

The performance of SED system is evaluated with poly-phonic sound event detection scores (PSDS)[21]. We compute PSDS using 50 operating points from 0.01 to 0.99. The PSDS parameters we used are shown in Table 2. Additionally, event-based macro F1 score (EB-F1) and segment-based macro F1 score (SB-F1) are used to analysis SED system. Event-based metrics are calculated based on the onset/offset of the predicted result. We set a 200 ms collar on onsets and a 200 ms / 20% of the events length collar on offsets. Segment-based metrics are calculated whether events are correctly predicted in each segment. we set a segment length to 1 second.
3.3. Experiment

3.3.1. Comparison of using pretrained weights and training models from Scratch

We conducted experiments to understand whether pretrained model is better than randomly initialized model. For pretrained model, the Efficientnet-B0 is initialized with pretrained weights and the rest of the network is initialized randomly. For training model from scratch, the whole network is initialized randomly. The results of this experiment are shown in Table 3. Note that we used weighted pooling strategy, mean teacher and mixup in this experiment. Experimental results show that the image pretrained weight can significantly improve the performance of the network on small audio dataset such as DCASE2021 Task4 dataset.

Table 3. Comparison of PSDS score when using pretrained weights and random weights

|                  | Pretrained | Random |
|------------------|------------|--------|
| PSDS score       | 0.62       | 0.50   |
| EB-F1            | 43.5%      | 38.0%  |
| SB-F1            | 72.1%      | 64.3%  |

3.3.2. Comparison of weighted pooling method

We conducted experiments on three pooling methods to investigate the influence of pooling method on weakly supervised learning. In addition to our proposed temporary attention mechanism and weighted pooling strategy, we also introduce average pooling as a comparison. We used pretrained weights, mean teacher and mixup in this experiment. The results of this experiment are shown in Table 4. The PSDS score shows that the performance difference between the two methods is small, but both of them are better than average pooling.

Table 4. Comparison of weighted pooling

|                  | PSDS score | EB-F1  | SB-F1  |
|------------------|------------|--------|--------|
| mean pooling     | 0.57       | 40.6%  | 69.7%  |
| weighted pooling | 0.62       | 43.5%  | 72.1%  |
| temporary attention | 0.61    | 43.2%  | 72.6%  |

3.3.3. Comparison to DCASE2021 Task4 baseline

DCASE2021 Task4 baseline[22] trained a CRNN from scratch, which uses channel attention as pooling method and mean teacher as semi supervised learning method. Our best model achieves a PSDS score of 0.62 on the validation set, which outperforms baseline's PSDS score of 0.52. In order to compare the details of the two models, we compare the segment based macro F1 score of each type of sound event in Table 5. The results show that our model outperforms baseline in detecting most sound events. In particular, our model greatly improves the detection performance of dishes, which has the characteristics of short duration and relatively difficult to detect.

Table 5: Segment-based macro F1 score of baseline and our model

| event class | Baseline | our model |
|-------------|----------|-----------|
| Blender     | 63.7%    | 60.4%     |
| Frying      | 58.1%    | 61.4%     |
| Cat         | 62.1%    | 65.4%     |
| Dog         | 65.4%    | 67.3%     |
4. Conclusions

In this paper, we have described our weakly and semi-supervised SED method. Our method used pretrained neural network weights, weighted pooling and mean teacher technique. The experimental evaluation using the DCASE2021 Task4 dataset showed that the proposed method outperformed the baseline model, which demonstrated the effectiveness of image pretraining for SED and our proposed weighted pooling.

References

[1] C. Clavel, T. Ehrette, G. Richard. Events detection for an audio-based surveillance system [C]. IEEE International Conference on Multimedia and Expo, 2005: 1306-1309
[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, Reliable detection of audio events in highly noisy environments, Pattern Recognition Letters 65(2015) 22–28, https://doi.org/10.1016/j.patrec.2015.06.026.
[3] J. Salamon, J.P. Bello, Feature learning with deep scattering for urban sound analysis, in: 23rd European Signal Processing Conference, EUSIPCO 2015, 2015, https://doi.org/10.1109/EUSIPCO.2015.7362478.
[4] N. Takahashi, M. Gygli, B. Pfister and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection", 2016.
[5] Lu, Rui. “BIDIRECTIONAL GRU FOR SOUND EVENT DETECTION.” (2017).
[6] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection", IEEE/ACM TASLP, vol. 25, no. 6, pp. 1291-1303, 2017.
[7] S. Adavanne, P. Pertila and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network", ICASSP, pp. 771-775, 2017.
[8] Zhi-Hua Zhou, A brief introduction to weakly supervised learning, National Science Review, Volume 5, Issue 1, January 2018, Pages 44–53
[9] Carboneau, M. et al. “Multiple instance learning: A survey of problem characteristics and applications.” Pattern Recognit. 77 (2018): 329-353.
[10] Berthelot, David et al. “MixMatch: A Holistic Approach to Semi-Supervised Learning.” ArXiv abs/1905.02249 (2019): n. pag.
[11] Berthelot, David et al. “ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring.” ArXiv abs/1911.09785 (2019): n. pag.
[12] Y. Gong, Y.-A. Chung, and J. Glass, “PSLA: Improving audio event classification with pretraining, sampling, labeling, and aggregation,” arXiv preprint arXiv:2102.01243, 2021.
[13] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “ESResNet: Environmental sound classification based on visual domain models,” in ICPR, 2020.
[14] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking CNN models for audio classification,” arXiv preprint arXiv:2007.11154, 2020.
[15] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Workshop on Detection and Classification of Acoustic Scenes and Events. New York City, United States, October 2019. URL: https://hal.inria.fr/hal-02160855.
[16] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." International Conference on Machine Learning. PMLR, 2019.
[17] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[18] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[19] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv preprint arXiv:1703.01780, 2017.

[20] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.

[21] Bilen Ç, Ferroni G, Tuveri F, et al. A framework for the robust evaluation of sound event detection[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 61-65.

[22] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In Workshop on Detection and Classification of Acoustic Scenes and Events. New York City, United States, October 2019. URL: https://hal.inria.fr/hal-02160855.