FRAMECNN: A WEAKLY-SUPERVISED LEARNING FRAMEWORK FOR FRAME-WISE ACOUSTIC EVENT DETECTION AND CLASSIFICATION

Szu-Yu Chou[1,2], Jyh-Shing Roger Jang [1], Yi-Hsuan Yang[2]

[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
[2]Research Center for IT innovation, Academia Sinica, Taipei, Taiwan
{fearofchou, yang}@citi.sinica.edu.tw, jang@csie.ntu.edu.tw

ABSTRACT

In this paper, we describe our contribution to the challenge of detection and classification of acoustic scenes and events (DCASE2017). We propose *framCNN*, a novel weakly-supervised learning framework that improves the performance of convolutional neural network (CNN) for acoustic event detection by attending to details of each sound at various temporal levels. Most existing weakly-supervised frameworks replace fully-connected network with global average pooling after the final convolution layer. Such a method tends to identify only a few discriminative parts, leading to sub-optimal localization and classification accuracy. The key idea of our approach is to consciously classify the sound of each frame given by the corresponding label. The idea is general and can be applied to any network for achieving sound event detection and improving the performance of sound event classification. In acoustic scene classification (Task1), our approach obtained an average accuracy of 99.2% on the four-fold cross-validation for acoustic scene recognition, comparing to the provided baseline of 74.8%. In the large-scale weakly supervised sound event detection for smart cars (Task4), we obtained a F-score 53.8% for sound event audio tagging (subtask A), compared to the baseline of 19.8%, and a F-score 32.8% for sound event detection (subtask B), compared to the baseline of 11.4%.

Index Terms— Deep learning, convolutional neural network, weakly supervised learning

1. INTRODUCTION

Developing an automatic system of acoustic event detection and classification is important for many real-world applications. For example, for multimedia based on its audio content, we can perceive the sound scene where we are within (e.g. home or office) [1]; when being applied to security surveillance devices, we can recognize abnormal sound sources (e.g. screaming, shouting, gun-shots) [2, 3]. Compared with acoustic event classification, *acoustic event detection* [4, 5, 6] is a more challenging task, because the system has to discriminate not only identifying the appearing of the sounds but also localizing the positions of those sounds in time.

However, most prior works for acoustic event detection propose a fully-supervised model for training. Such methods usually require a training data set that contains the annotation of the temporal position of the acoustic events. The limitation of *strongly labeled data* is that they are difficult to collect. As a consequence, such a data set may only include a few classes of sound and small-scale sound . In comparison, the so-called *weakly labeled data* only requires annotations of the occurrence of the acoustic events at the clip level, not the frame level. Therefore, such weakly labeled data is easier to amass and therefore useful for large-scale industrial applications.

There has been a number of works to propose a weakly-supervised learning method for music auto-tagging [7], acoustic event detection [8, 9] and image localization and segmentation [10, 11, 12]. Compared to the fully-supervised setting, weakly-supervised learning only relies on the weakly labeled annotations data.

In this paper, we focus on weakly-supervised learning method for acoustic event detection. Our key idea is to consciously identify the sound for each frame, forcing the network to pay attention to details of the sound clip. We extend a recent CNN for acoustic event classification by adding a branch of network for predicting a acoustic event of each frame by training on only weak annotations data. This branching network uses a small fully convolutional network (FCN) to identify the appearing of the sounds in a frame-to-frame manner. Our network is simple to implement and train given any CNN structure on the fly, such as the residual network [13] or the inception network [14]. Additionally, the branch only requires a small computational overhead, making a network can achieve acoustic event detection and improve the performance for acoustic event classification.

2. PROPOSED METHOD

2.1. Data processing

We use the 128-bin log mel-spectrogram as the audio feature to the neural network, which has been used widely in the literature [7, 9, 15]. The mel-spectrograms are computed by short-time Fourier transform with 2,048-sample, quarter-overlapping windows, for audio sampled at 44.1kHz. The mel-scale is to reduce the dimensionality along the frequency axis. The feature extraction was computed using the librosa library [1] [16]. After the extraction process, we standardize by removing mean and divided by standard deviation derived from training set.

2.2. Network Architecture

For the acoustic event classification, we use the CNN structure proposed by Oord [15]. The input of network is a mel-spectrograms, with 128 frequency bins and 862 frames. The first convolutional layer contains 256 kernels with size 128×4, with max pooling 1×4 and zero-padding 1×2. The second and third convolutional layer contains 512 kernels with size 1×4, with max pooling 1×4 and zero-padding 1×2. After the last convolutional layer, similar with

[1]https://github.com/librosa

Oord [15], we apply a global temporal pooling layer across the entire time axis. In the global temporal pooling layer, we concatenate the output of three pooling methods the mean, the maximum and the variance, leading to an output size of 1 536 for this layer. Finally, we use two fully-connected layers with 1,024 neurons for learning high-level representation and use the learnt high-level representation to classify the acoustic event.

For the acoustic event detection, we add a branch on the same CNN structure after the last convolutional layer. In the branch, we employ transpose convolution network to reconstruct the original audio of activations and perform frame-wise classification by training on weakly annotation data. The first transpose convolutional layer contains 512 kernels with size 1×4, with stride 1×4. The second transpose convolutional layer contains 256 kernels with size 1×4, with stride 1×4. Finally, we use convolutional layer containing the number of class kernels with size 1×1 to predict the frame-wise class label. While training, we assume that all the frames of a clip have the same class label as the clip.

Note that the global temporal pooling layer and the FCN structure allows the model to process audio clips of variable length for acoustic event detection and classification.

3. EVALUATION

3.1. Acoustic scene classification

In this task, we compare the provided baseline results. The baseline system is a deep neural network using log mel-band energies with 5 context frames as features. The feature is extracted with a frame size of 40ms and 20ms hop size. The network contains two fully-connected layers with 50 neurons and uses dropout technique to avoid overfitting. For the classification task, the decision layer uses the softmax function as output activation function. Table 1 shows the average classification accuracy over 4 evaluation folds for acoustic scene classification. As can been seen, our network obtains an accuracy of 99.2%, compared to the average baseline of 74.8%. The result demonstrates our network significantly outperforms the baseline system. On the other hand, we found our weakly-supervised framework can not only perform frame-wise classification but also improve the performance of classification accuracy. Note that we did not use any output of frame-wise classification to make the clip-level decision.

3.2. Large-scale weakly supervised sound event detection for smart cars

The evaluation of sound event detection for smart cars are divided into two subtasks. The first task (subtask A) considers the sound event detection without timestamps, which is similar to acoustic multi-lable classification (i.e. we only need to identify the appearing of sound in audio). The evaluation metrics is based on class-level F1-score. The second task (subtask B), on the other hand, considers the sound event detection with timestamps, which means that we need to identify a boundary for the appearing of sound in audio. This subtask considers two evaluation metrics: segment-based error rate and segment-based F1-score, using one-second segments. The model and feature extraction of baseline system is the same with the one for acoustic scene classification, but it uses the sigmoid function as output activation function.

Table 2 compares three metrics F1-score, precision and recall with class-level of subtask A for sound event detection without timestamps. As we can see, the proposed method achieves an F1-score of 50.0%, compared to the baseline of 13.1%. The result shows our network performs remarkably better than the baseline system. On the other hand, the result of subtask B is shown in Table 3. Again, our model also can perform well by given the boundary of time. As we can see, the value of F1-score metric of the subtask B is much lower than the subtask A, which demonstrates the difficulty of identifying the boundary of the sound in audio by training the model with only the weak annotation.

Table 1: Experimental result for acoustic scene classification accuracy, averaged over 4 evaluation folds

Acoustic Scene	Baseline	Proposed Method
Beach	75.3%	**98.6%**
Bus	71.8%	**100.0%**
Cafe / Restaurant	57.7%	**100.0%**
Car	97.1%	**98.8%**
City center	90.7%	**99.3%**
Forest path	79.5%	**98.7%**
Grocery store	58.7%	**99.6%**
Home	68.6%	**98.1%**
Library	57.1%	**99.0%**
Metro station	91.7%	**99.3%**
Office	99.7%	99.7%
Park	70.2%	**99.0%**
Residential area	64.1%	**98.2%**
Train	58.0%	**98.9%**
Tram	81.7%	**100.0%**
Overall accuracy	74.8%	**99.2%**

Table 2: Experimental result of subtask A for sound event detection without timestamps

Class-based	Baseline	frameCNN
F1-score	13.1%	**53.8%**
Precision	12.2%	**54.0%**
Recall	14.1%	**55.4%**

4. CONCLUSION

In this paper, we presented a novel weakly-supervised framework to enforce the deep learning network by discriminating to meticulous detail of each sound. Extending a recent convolutional neural network for acoustic event classification, we add the branch of network for directly predicting a acoustic event of each frames by training on only weak annotations data. Our experiment on three tasks shows that our network outperforms the baseline system in acoustic scene classification and weakly supervised sound event detection. Moreover, The proposed network can be added on any CNN structure,

Table 3: Experimental result of subtask B for sound event detection with timestamps

Segment-based metric	Baseline	frameCNN
Error rate (ER)	1.02	**0.86**
F1-score	13.8%	**32.8%**

and only requires substantially fewer parameters and less computation to achieve state-of-the-art performances.

5. REFERENCES

[1] J. C. Wang, C. H. Lin, B. W. Chen, and M. K. Tsai, "Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 607–613, 2014.

[2] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 6460–6464.

[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] T. Fischer, J. Schneider, and W. Stork, "Classification of breath and snore sounds using audio data recorded with smartphones in the home environment," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 226–230.

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2015, pp. 1–6.

[6] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 171–175.

[7] J.-Y. Liu and Y.-H. Yang, "Event localization in music auto-tagging," in *Proceedings of the ACM on Multimedia Conference*. ACM, 2016, pp. 1048–1057.

[8] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the ACM on Multimedia Conference*, 2016, pp. 1038–1047.

[9] T. W. Su, J. Y. Liu, and Y. H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *Proceedin of International Conference on Acoustics, Speech and Signal Processing*, March 2017, pp. 791–795.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[11] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[13] 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*, June 2016, pp. 770–778.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedin gs of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2013, pp. 2643–2651.

[16] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nietok, "librosa: Audio and music signal analysis in python," in *annual Scientific Computing with Python conference*, 2015.