FREQUENCY DEPENDENT SOUND EVENT DETECTION FOR DCASE 2022 CHALLENGE TASK 4

Technical Report

Hyeonuk Nam, Seong-Hu Kim, Deokki Min, Byeong-Yun Ko, Seung-Deok Choi, Yong-Hwa Park

Korea Advanced Institute of Science and Technology
Department of Mechanical Engineering, 291 Daehak-ro,
Yuseong-gu, Daejeon 34141, South Korea
{frednam, seonghu.kim, minducky, b.y.ko, haroldchoi6, yhpark}@kaist.ac.kr

ABSTRACT

While many deep learning methods on other domains have been applied to sound event detection (SED), differences between original domains of the methods and SED have not been appropriately considered so far. As SED uses audio data with two dimensions (time and frequency) for input, thorough comprehension on these two dimensions is essential for application of methods from other domains on SED. Previous works proved that methods those address on frequency dimension are especially powerful in SED. By applying FilterAugment and frequency dynamic convolution those are frequency dependent methods proposed to enhance SED performance, our submitted models achieved best PSDS_1 of 0.4704 and best PSDS_2 of 0.8224.

Index Terms— Sound Event Detection, FilterAugment, Frequency Dynamic Convolution

1. INTRODUCTION

Sound event detection (SED) which aims to classify desired sound event classes and their time localization (onset and offset) in a given audio signal has been rapidly growing with advancement of deep learning (DL) methods [1, 2, 3, 4, 5, 6, 7]. As 1D audio data with time dimension is usually expanded into 2D data with time and frequency dimension for audio signal processing, 2D time-frequency audio data are usually used for DL based SED by treating 2D audio data as 2D image data and applying DL methods for image data [6, 7, 8]. Although DL methods for 2D image data showed powerful performance on their own domain, they have inherent inconsistency on SED which arises from the difference between 2D image data and 2D audio data. While 2D image data consists of two same dimensions representing the same physical quantity (location), 2D audio data consists of two different dimensions representing different physical quantity (time and frequency). Considering that time is somewhat similar to location as they both are translation equivariant (certain pattern is still the same entity when it is moved along location or time dimensions) while frequency is not translation equivariant because each frequency value represents different characteristics from the others, frequency is the dimension to be thoroughly considered in SED [6, 7]. In this work, we especially apply methods that address such difference between 2D image data and 2D audio data for SED by addressing issues of frequency dimension in 2D audio data. SED models illustrated in this report could be trained using code available in GitHub[1]. This repository includes both FilterAugment and frequency dynamic convolution.

2. METHODS

2.1. FilterAugment

FilterAugment is proposed to generalize SED model to various acoustic environments, to make SED model more like human who can classify sound events from different acoustic environments [6]. By randomly dividing frequency ranges into several frequency bands and randomly applying weights on the frequency bands of a Mel spectrogram, FilterAugment could approximately simulate acoustic environments that results in different frequency weights on different frequency bands. Although resulting Mel spectrogram might sound unnatural, it is simple to use and effective on SED as shown in [6]. FilterAugment could emphasize different frequency bands of the same data every epoch, thus it helps to train SED model to recognize time-frequency patterns from wider frequency ranges. Without FilterAugment, SED model might be trained to recognize patterns from most distinct time-frequency patterns instead. Code for SED with FilterAugment is available in GitHub[2].

There are two types of FilterAugment: step and linear type. Step type FilterAugment applies constant weights over each frequency bands and the weights change abruptly across the boundary of frequency bands, while linear type FilterAugment applies continuous weights over frequency bands by assigning weights on frequency boundaries and then linearly interpolating weights between the boundaries. In this work, these different types of FilterAugment ensemble averaged to increase variety of SED model capacity thus enhance performance of ensemble aveaged model.

Hyperparameter setting used in this work is as follows: step type FilterAugment with dB range = (-4.5, 6), band number range = (2, 5) and minimum bandwidth = 4, linear type FilterAugment with dB range = (-6, 4.5), band number range = (3, 6) and minimum bandwidth = 7.

This work was supported by "Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20204030200050)

[1] https://github.com/frednam93/FDY-SED

[2] https://github.com/frednam93/FilterAugSED

Table 1: Training settings with their best PSDS_1, PSDS_2 scores and collar-based F1 score.

Setting Index	Seed	FilterAugment Type	Attention Dimension	PSDS_1	PSDS_2	CB-F1
1	21	step	class	0.4446	0.6780	0.536
2	42	step	class	0.4554	0.6719	0.536
3	42	linear	class	0.4510	0.6702	0.536
4	42	step	time	0.4548	0.6743	0.533

2.2. Frequency Dynamic Convolution

Frequency dynamic convolution is proposed to weaken translation equivariance of 2D CNN on frequency axis and to improve CNN kernel's adaptability to the input at the same time [7]. It is inspired by temporal dynamic models that applied dynamic convolution proposed for image recognition into speaker verification [9, 10, 11]. As different frequency regions exhibit different frequency patterns of sound events, different convolution kernels should be used on different frequency regions. Thus, frequency dynamic convolution applies convolution kernel that dynamically adapts to each frequency bin of the convolution input. By applying adaptive kernels that differs on each frequency bin, translation equivariance is weakened in frequency dynamic convolution and it helps recognizing more complex time-frequency pattern for SED as shown in analysis of class-wise comparison [7]. In this work, 4 basis kernels and temperature of 45 are used for frequency dynamic convolution.

2.3. Implementation Details

The code used to train SED model submitted for this participation could be found in Github repository mentioned in introduction. It is derived from DCASE 2021 Challenge Task 4 baseline [2, 3]. Input audio data with length of 10 seconds and sampling rate of 16 kHz are used. They are converted to log Mel spectrogram with number of FFT, hop length and number of Mel bins as 2048, 256 and 128 respectively. Each batch of log Mel spectrograms are normalized to be between 0 and 1 on batch and time dimensions. For data augmentation, frame shift, mixup [12], time masking [13] and FilterAugment [6] are used. To utilize unlabeled dataset, mean teacher method is used [14].

The model has CRNN architecture composed of 7 CNN layers and 2 BiGRU layers, then frame-wise fully connected (FC) layer makes strong prediction of the input [15]. Each CNN layer is composed of convolution module, batch normalization, context gating, dropout, and then 2D average pooling. The first CNN layer uses normal 2D convolution and the rest six CNN layers use frequency dynamic convolution [7]. After 2 biGRU layers, frame-wise FC layer outputs strong prediction while other FC layer followed by Softmax extracts attention weights to apply on strong prediction to result in weak prediction. On the output, weak prediction masking or weak SED [5] is applied and then applied by median filter. Different length of median filter is applied on each sound event class: 5 for alarm/bell ringing, cat, dish, dog and speech and 11, 67, 61, 49, 17 for blender, electric shaver/toothbrush, frying, running water and vacuum cleaner respectively.

Evaluation metrics used in this report are polyphonic sound detection score (PSDS) [16] and macro event-based F1 score [4]. PSDS_1 and PSDS_2 are the metrics used in DCASE 2022 challenge Task4 [3], which favors SED system that predicts accurate timestamp and SED system that does not produce cross triggers respectively.

Table 2: Performance of submissions with number of models ensemble averaged.

Submit Index	# Models	PSDS_1	PSDS_2	CB-F1
1	31	**0.4704**	0.6866	0.543
2	12	0.4703	0.7002	0.541
3	53	0.0606	**0.8224**	0.199
4	150	0.0584	0.8195	0.536

2.4. Ensemble Averaging

Table 1 shows four settings used in this work with their best PSDS scores and collar-based F1 score (CB-F1) of single SED model in each setting. For each setting, 48 training runs are separately done resulting in 96 models (48 student models and 48 teacher models). The settings differ by seed, FilterAugment type, and attention type for weak prediction pooling. The seed is used are 21 and 42. FilterAugment types used are step and linear as illustrated in 2.1. Attention types are class and time, which means the dimension on which Softmax applied to obtain attention weights. Using different settings resulted in more diverse SED models thus resulted in better ensemble averaged models.

3. RESULTS

Table 2 shows performance of submitted models which are ensemble averaged model of 4 settings in Table 1. We chose best 31 student and teacher models on PSDS_1 for submission 1, best 12 student models on PSDS_1 for submission 2, best 53 student models on PSDS_2 for submission 3 and best 150 student and teacher models on PSDS_2 for submission 4. For submission 1 and 2, weak prediction masking is applied and for submission 3 and 4, weak SED is applied [5]. As a result, the best PSDS_1 score is .0.4704 and the best PSDS_2 score is 0.8224.

4. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[2] N. Turpault, R. Serizel, A. Parag Shah, and J. 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*, 2019.

[3] DCASE. Dcase 2022 challenge task4: Sound event detection in domestic environments. [Online]. Available: https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments

[4] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[5] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," DCASE2021 Challenge, Tech. Rep., 2021.

[6] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[7] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint arXiv:2203.15296*, 2022.

[8] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, "Deep learning based cough detection camera using enhanced features," *Expert Systems with Applications*, vol. 206, 2022.

[9] S.-H. Kim and Y.-H. Park, "Adaptive convolutional neural network for text-independent speaker recognition," in *Proc. Interspeech*, 2021, pp. 641–645.

[10] S.-H. Kim, H. Nam, and Y.-H. Park, "Temporal dynamic convolutional neural network for text-independent speaker verification and phonemetic analysis," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[11] S. H. Kim, H. Nam, and Y. H. Park, "Decomposed temporal dynamic cnn: Efficient time-adaptive network for text-independent speaker verification explained with speaker activation map," *arXiv preprint arXiv:2203.15277*, 2022.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[13] 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," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[14] 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*, vol. 30, 2017.

[15] E. Çakır, 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.

[16] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 61–65.