Sound event detection of domestic activities using frequency dynamic convolution and BEATs embeddings

Technical Report

Grigorios-Aris Cheimariotis[*]

Democritus University of Thrace
Dpt. of Electrical and Computer Engineering
Xanthi 67100, Greece
gcheimar@ee.duth.gr

Nikolaos Mitianoudis[*]

Democritus University of Thrace
Dpt. of Electric and Computer Engineering
Xanthi 67100, Greece
nmitiano@ee.duth.gr

ABSTRACT

This technical report describes one submission for Dcase2023 Task 4a "Sound event detection of domestic activities". The methodologies proposed are based on the baseline system, which is provided by the organizers, and consist mainly of feature extraction by passing spectrograms through frequency dynamic convolution network, concatenation of these features with BEATS embeddings, use of BiGRU for sequence modelling. Also, a mean-teacher model is employed. The results for the submissions, when using audioset real strong-labelled data are: PSDS1 0.496 PSDS2 0.788 and when the aforementioned data subset is not used are: PSDS1 0.516 PSDS2 0.781.

Index Terms— Sound event detection, BiGRU, mean-teacher model, frequency dynamic convolution, embeddings

1. INTRODUCTION

Sound event detection of domestic activities is of specific interest for various applications, including assisting autonomous living of elderly. Through the years the dataset provided by dcase community has been enhanced. It includes strong labeled, weak labeled and unlabeled clips of 10 seconds, in which 10 classes of events may occur for a specific duration. Although some classes are different from what is of interest, in assisting autonomous living, it is highly probable that an efficient system of detecting events, will be efficient in detecting additional events. Various deep learning models of different complexity have been proposed the last years and achieve a high performance in terms of different metrics. These models, in most cases, have as input a two-dimensional spectrogram which represent the clips of 10 seconds. The spectrograms are processed with image processing models (e.g convolutional neural networks), which may be altered to better capture the nature of sound spectrograms. One such successful adaptation to sound spectrograms is frequency dynamic convolutional network [1]. Complementary to spectrograms, embeddings of pretrained models are also used. These embeddings may occur by training in more audio events (audioset) or images or video. Various embeddings enhanced the performance of models. Lately, BEATS (audio pre-training with acoustic tokenizers) embeddings achieve significantly better performance in various tasks e.g. audio classification [2]. The sequence modeling is attempted by other bi-directional GRU or transformers. Mean teacher student is proven efficient in leveraging unlabeled data [3].

The proposed model adopts and modifies the baseline (version 2023) [4] by employing specific versions of the aforementioned modules. It combines BEATS embeddings with the output of a frequency dynamic convolution network which processes the spectrogram of 10 sec clip. The combined embeddings pass through either a BiGRU (bi-directional recurrent unit) and finally a classification module. The model that consists of the above modules (student), has an identical teacher model which is updated with weighted of average of the weights of student.

2. DATASET

The dataset of DCASE 2023 is composed by: labeled training set, unlabeled in domain training set and synthetic set with strong annotations. The audio clips are sampled at 44,100 Hz and have duration of 10 seconds at maximum. Each audio clip contains at least one sound corresponding to one of the 10 possible classes.

3. SPECTROGRAM AND DATA AUGMENTATION

Spectrograms are produced as in the baseline method and mixup [5] with soft labels is randomly applied with a probability 0.5. Specifically, audio clips are resampled at 22,050 Hz and log mel-spectrogram are extracted from them. The size of the analysis window is 2048, the hop length is 256 and the number of mels is chosen to be 128.

4. FREQUNCY DYNAMIC CONVOLUTIONAL NETWORK

Frequency-adaptive convolution has been already proposed for sound event detection and outperforms various other methods for DESED task, as indicated by polyphonic sound detection score. Frequency dynamic convolution is applied to downplay the effect of translation equivariance which is not desirable due to the

[*] Thanks to the project "Improvement of the Quality of Life and Activity for the Elderly" (MIS 5047294)

specific nature of sound spectrograms. In order to combine a frequency-dynamic convolution network with resource-demanding transformer, a miniaturized (half-scale of original) frequency dynamic convolutional was tested. Lower performance indicated that a wider convolution network (double-scale) may perform better when combined by BiGRU light-weight sequence modelling. However, no improvement was achieved. Therefore, the proposed Frequency dynamic convolutional network is very similar with the original one proposed by authors [1]. Its properties are: kernel size 3 in all layers, number of filters in each layer [32,64,128,256,256,256,256], activations are context-gating[6][7], 0.5 dropout is used.

5. BEATS EMBEDDINGS

BEATs embeddings are proposed by DCASE23 organizers as a state-of-the art method for pretrained embeddings [2]. BEATs are extracted from a Bidirectional encoder which learns representations form audio transformers, where a tokenizer and an audio self-supervised model are optimized. The BEATs embeddings are concatenated with the output of Frequency dynamic CNN with the pool1d function of baseline. Specifically, the embeddings are reshaped to match the time dimension of frequency dynamic CNN by averaging pooling and then they are concatenated so as the time frames coincide.

6. BIGRU

Bidirectional GRU is as used in the baseline with number of RNN cells 128 but the number of layers is modified to 3 from 2 in the baseline. No dropout was used within the RNN but a dropout of 0.5 is applied to the output of RNN. Then, the classification module (including one dense layer) is used with attention and it is unmodified from baseline.

7. MEAN TEACHER

Mean teacher method was employed without modifications from baseline.

8. TRAINING DETAILS

The training epochs' number is set for 400 epochs with a rampup of 50 epochs and a patience of 100 epochs. The target learning rate is 0.001 and the optimizer is Adam (as in the baseline). Validation objective function was the sum of PSDS1 and PSDS2 (Polyphonic sound event detection score) as firstly defined in [8] but with a newer implementation [9][10]. The training batch size is set to [24(synth),24(weak),48(unlabeled)] and the validation batch size is set to 64. Weak split (the percentage of samples used for training) was set to 98%. An NVIDIA RTX A6000 GPU was used and the total number of trainable parameters are 3.3 M, the model size is 26,447MB and multiply-accumulate operations (MACs) are 3.497G..

9. RESULTS

The results are reported as instructed by DCASE guidelines with average of a 3-fold identical repetition of each model, to account for random weight initialization and are shown in Table 1. Although, PSDS1 score is close to the baseline method, PSDS2 (0.788) score is higher than the baseline (0.762). However, when trained without real strong data PSDS1 is enhanced to 0.516, which is higher of the best baseline score 0.5. When training stopped at 313 and 322 epochs, the training time was about 8h. Energy consumption extracted from codecarbon for 3 runs with the same hyperparameters and seed number, had fluctuations 1.49+0.13, 1.95+0.043, 1.09+0.035 kwh (training+testing energy).

Table 1 Results (average and standard deviation of 3 runs with the same hyperparameters). For energy consumption only average is shown here.

	PSDS1	PSDS2	Event-f1-macro	Intersection f1-macro	Energy consumption(train+test)
Duth (audioset strong)	0.496+-0.007	0.788+-0.005	0.595 +- 0.005	0.811+-0.001	1,51+0.03 kWh
Duth (w/o audioset strong)	0.516+-0.002	0.781+-0.008	0.597+-0.006	0.818+-0.001	1.783+0.34 kWh

10. CONCLUSION

Although the method proposed, adopts baseline method and uses modules tested before with small modifications, it achieves higher psds scores than what was reported in dcase2022, without ensemble and with a comparatively light-weight system.

11. ACKNOWLEDGMENT

This study was made possible through the project "Improvement of the Quality of Life and Activity for the Elderly" (MIS 5047294) which was implemented under the "Support for Regional Excellence" program, financed by the "Competitiveness, Entrepreneurship and Innovation" program (NSRF 2014-2020) and funded jointly by Greece and the European Union (European Regional Development Fund).

12. REFERENCES

[1] H. Nam, S. H. Kim, B. Y. Ko, and Y. H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, pp. 2763–2767, 2022.

[2] S. Chen *et al.*, "BEATs: Audio Pre-Training with Acoustic Tokenizers," 2022.

[3] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 1196–1205, 2017.

[4] L. Delphin-Poulat and C. Plapous, "Mean Teacher With Data Augmentation for Dcase 2019 Task 4," *Dcase*, pp. 2018–2020, 2019.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–13, 2018.

[6] L. JiaKai, "Mean Teacher Convolution System for DCASE 2018 Task 4," *Detect. Classif. Acoust. Scenes Events 2018*, no. November, 2018.

[7] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with Context Gating for video classification," pp. 1–8, 2017.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, 2016.

[9] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 61–65, 2020.

[10] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold Independent Evaluation of Sound Event Detection Scores," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 1021–1025, 2022.