SOUND EVENT DETECTION SYSTEM FOR DCASE 2021 CHALLENGE

Technical Report

Jakub Bajzik

University of Zilina
Department of Mechatronics and Electronics
Žilina 010 26, Slovak Republic
jakub.bajzik@feit.uniza.sk

ABSTRACT

This paper presents the systems proposal for the DCASE 2021 challenge Task 4 (Sound event detection and separation in domestic environments). The aim is to provide the event time localization timestamps in addition to event class probabilities. In this paper, the two systems are proposed. System 1 is a convolutional neural network trained for sound event classification using only weakly labeled and unlabeled data. The strong labels are obtained using the class activation mapping technique. System 1 does not reach the baseline performance. System 2 is the convolutional neural network and recurrent neural network which uses the class activation mapping technique as a part of the attention mechanism to increase the baseline performance. The second model was trained using weakly labeled, strongly labeled, and unlabeled data. Both architectures are based on the Mean Teacher baseline system 2021.

Index Terms— Semi-supervised learning, sound event detection, class activation maps, attention

1. INTRODUCTION

DCASE 2021 task 4 targets on the sound events detection using weakly labeled data, unlabeled data (without timestamps), and synthetic data. The systems should provide the event time localization (timestamps) in addition to the event class probabilities. The challenge encourages the participants to explore the possibility to exploit a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set. The challenge allows participate in three subtasks [1]:

1. SED without sound separation
2. SED with sound separation
3. Sound separation

I participate only in the first subtask, so the proposed system performs only SED without sound separation. The target is to explore the possibility to use the technique called class activation mapping (CAM) [2]. I propose two systems (the challenge allows submit 4 systems), which both are based on baseline Mean Teacher architecture [3] and both use the CAM technique but in different ways.

System 1 is a convolutional neural network (CNN) trained for sound event classification using only weakly labeled and unlabeled data. The strong labels are obtained in test steps using the CAM technique.

System 2 is CNN with RNN which uses the CAM technique as a part of the attention mechanism to increase the baseline performance. The second model was trained using weakly labeled, strongly labeled, and unlabeled data.

The performance of system 1 is significantly lower than the baseline performance, but indicates the possibility of usage for attending the most class-significant frames in the audio signal. The baseline was outperformed by system 2, where the CAM technique is the part of the attention mechanism.

2. RELATED WORK

The previous research in the field shows the benefits of convolutional recurrent neural networks (CRNN) for SED tasks. The challenge is to exploit a large amount of unbalanced and unlabeled data together with a small weakly annotated dataset. The semi-supervised approach was introduced in [4], where the two models are trained separately. The second model is trained using unlabeled data with pseudo labels obtained as the predictions of the first model. In [5], the Mean Teacher model that averages model weights was introduced. The work [6] proposes the SED system with context gating CNN and RNN followed by softmax attention on weak predictions. The semi-supervised model uses the weakly labeled data and the maximize use of unlabeled data. The detailed analysis of DCASE2020 task 4 sound event detection baseline is introduced in [7]. The baseline implementation of the Mean Teacher model is based on the work [5] and the architecture is inspired by work [6]. The systems proposed in this work are based on the baseline system. This work aims to investigate the usage of class activation maps [2] for semi-supervised audio detection. A similar approach was used in study [8], where authors proposed also a Time-Frequency audio segmentation.

3. MATERIALS AND METHODS

3.1. Dataset

The development dataset for SED is composed of 10-sec soundscapes recorded in a domestic environment or synthesized to simulate a domestic environment [3]. The data are provided in 3 different splits: weakly labeled training set (1578 clips), unlabeled training set (14412 clips) and synthetic set with strong annotations (10000) [1]. The proposed systems were trained using different subsets as described in the Table 1.

System	Description	Subsets
System 1	CNN - CAM	weak, unlabeled
System 2	CRNN - CAM attention	strong, weak, unlabeled

Table 1: Subsets used for training the proposed systems.

3.2. Audio processing

The proposed systems work with 16,000 Hz sampling frequency. The Log-Mel spectrograms are extracted from the 10-seconds long audio clips using a 2048 window size and 128 bins. The clips are padded, if the length is less than 10-seconds.

3.3. Class activation maps

The class activation maps for a particular class allows simply localize the discriminative image regions used by the CNN to identify the class. It was introduced as the weakly supervised object localization method in [2]. The advantage of global average pooling (GAP) is that the network can retain its localization ability until the final layer.

In the SED task, we are working with 2-dimensional Log-Mel Spectrograms, which are pooled by CNN in the frequency domain. The features are frame vectors in each channel. Therefore, the CAM in this work is computed on the feature vector, not the feature map. The class activation vector (CAV) for particular class c is computed as

$$V_c(i) = \sum_n w_n^c f_n(i) \tag{1}$$

where i is the frame number, n is the channel number, f_n is the feature vector and w_n^c is the weights vector of the particular class neuron.

4. SYSTEMS DESCRIPTION

4.1. System 1

System 1 diagram is depicted in Fig. 1. The Log-Mel Spectrogram is pooled over the frequency domain to obtain a single frame vector in each convolutional channel. The output classification dense layer is fully connected with the GAP layer. As multiple events can be present in an audio recording, the sigmoid activation is used. The vector class activation vector V for particular class c is computed using 1. Strong predictions are obtained from V after sigmoid activation and multiplication with weak predictions of each class.

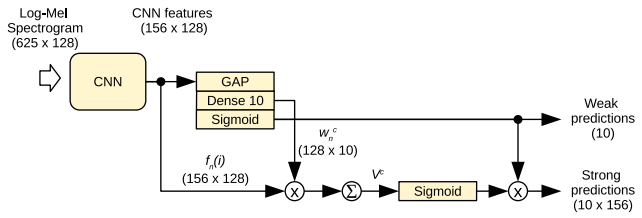

Figure 1: Diagram of the proposed system 1.

The system is the Mean Teacher model, where the consistency cost and classification cost are calculated only on clip level, so there are no strong labels needed in the training step. The model was trained using weak and unlabeled data for classification. The teacher's model weights are the moving average of the student's weights.

4.2. System 2

The system is the Mean Teacher model, which uses the consistency cost and classification cost on clip and frame levels in the same way as the baseline system. The model was trained using synthetic strongly labeled data, weakly labeled data, and unlabeled data.

The proposed mechanism for attending the strong predictions using CAM is depicted in Fig. 2. Class activation vector V and dense layer output is the same size, so they can be added together just before the sigmoid activation.

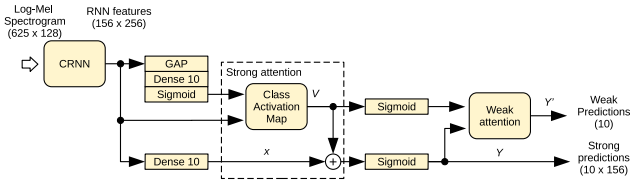

Figure 2: Diagram of the proposed system 2.

The attention from baseline [6] is partially retained. To obtain the weak predictions, the strong predictions are combined with class activation vectors. The attention on weak predictions is defined as

$$Y' = \frac{\sum_i \sigma(V) \odot \sigma(x+V)}{\sum_i \sigma(V)} \tag{2}$$

where i is the frame number and $\odot$ denotes the element-wise multiplication. Strong predictions are defined as

$$Y(i) = \sigma(x+V) \tag{3}$$

where σ is the sigmoid activation, x is the dense layer output and V is the class activation vector obtained by 1.

5. RESULTS

In the test step, the teacher output is more likely to be correct but principally both model outputs can be used for predictions. The performance of systems in Table 2 is evaluated with polyphonic sound event detection scores (PSDS) [9], macro-averaged event-based and intersection-based F1 score computed over recordings in the development validation set. The results are for teacher models. PSDS is calculated for two different scenarios that emphasize different systems properties, as described in [1].

System	PSDS 1	PSDS 2	$F1_{intersec}$ [%]	$F1_{event}$ [%]
Baseline	0.353	0.553	79.5	**42.1**
System 1	0.165	0.348	75.2	14.1
System 2	**0.374**	**0.586**	**81.3**	41.2

Table 2: Comparison of models on development validation set.

6. CONCLUSION

The results show, that the performance of system 1 with a single CNN is relatively low. It can suffer from the absence of recurrent network and strongly labeled synthetic data. The experiments showed that the key point is the normalization of the class activation vector. The non-linear normalization unit is the sigmoid activation, but it can be replaced with some sophisticated algorithm to achieve better results in the future.

System 2 outperforms the baseline system slightly. It outlines the utility of the proposed attention mechanism. There is still a place for improvement, so the proposed attention mechanism can be redesigned in future work.

7. REFERENCES

[1] "DCASE." [Online]. Available: http://dcase.community/

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *arXiv:1512.04150 [cs]*, Dec. 2015, arXiv: 1512.04150. [Online]. Available: http://arxiv.org/abs/1512.04150

[3] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019, pp. 253–257. [Online]. Available: http://hdl.handle.net/2451/60771

[4] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," *arXiv:1807.10501 [cs, eess]*, July 2018, arXiv: 1807.10501. [Online]. Available: http://arxiv.org/abs/1807.10501

[5] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv:1703.01780 [cs, stat]*, Apr. 2018, arXiv: 1703.01780. [Online]. Available: http://arxiv.org/abs/1703.01780

[6] L. Jiakai, "Mean Teacher Convolution System for DCASE 2018 Task 4," Tech. Rep., 2018.

[7] N. Turpault and R. Serizel, "Training Sound Event Detection On A Heterogeneous Dataset," *arXiv:2007.03931 [cs, eess]*, July 2020, arXiv: 2007.03931. [Online]. Available: http://arxiv.org/abs/2007.03931

[8] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound Event Detection and Time-Frequency Segmentation from Weakly Labelled Data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, Apr. 2019, arXiv: 1804.04715. [Online]. Available: http://arxiv.org/abs/1804.04715

[9] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," *arXiv:1910.08440 [cs, eess]*, Feb. 2020, arXiv: 1910.08440. [Online]. Available: http://arxiv.org/abs/1910.08440