Sound event detection in domestic environments with weakly labeled data and soundscape synthesis

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, Justin Salamon

SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS WITH WEAKLY LABELED DATA AND SOUNDSCAPE SYNTHESIS

Nicolas Turpault[1], *Romain Serizel*[1], *Ankit Parag Shah*[2], *Justin Salamon*[3]

[1]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[2]Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, United States
[3]Adobe Research, San Francisco CA, United States

ABSTRACT

This paper presents task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge and provides a first analysis of the challenge results. The task is a follow up to task 4 of DCASE 2018, and involves training systems for large-scale detection of sound events using a combination of weakly labeled data, i.e. training labels without time boundaries, and synthesized strongly-labeled data. The paper focuses in particular on the additional synthetic, strongly labeled, dataset provided this year. More details about the analysis of the results will be provided after the evaluation period (July 2019).

Index Terms— Sound event detection, Weakly labeled data, Semi-supervised learning, Synthetic data

1. INTRODUCTION

Sounds carry a large amount of information in our every day life and we depend upon sounds to better understand the changes in our physical environment and perceive the events surrounding us. We perceive the scene (eg. in an airport, in a house etc) as well as individual sound events (eg. car honking, footsteps, speech etc). Sound event detection within an audio segment refers to the task of classifying the sound events in the category as well as temporally locating the occurrences of an event in the audio recording. Sound event detection has potential applications in context awareness (noise monitoring in smart cities) [1, 2], surveillance [3], urban planning [1], multimedia information retrieval [4, 5] , domestic applications such as smart homes, health monitoring systems, home security solutions [6, 7, 8] to name a few and hence the field has gathered interest in the broader areas of machine learning and audio processing.

Sound Event detection using weak label learning has gathered a lot of interest [6, 9, 10, 11] in the community as it address problems in developing approaches dependent on strongly labeled data. For instance, strongly labeled data is time consuming and difficult to annotate as it requires annotation of temporal occurrence as well as presence of absence of an sound event. Additionally, strongly labeled data annotations has the probability of introducing human error given the ambiguity of the onset and offset interpretation for a sound event. In case of the weakly labeled data, we only have information about whether an event is present in the recording or not. We have no information about occurrences or information about the temporal locations of a given sound event in an audio clip with weakly annotated data. In real world datasets, it is critical to build systems which generalizes over a large number of classes, variety of distribution of audio events. For such cases, gathering weakly labeled data may be feasible to collect in large quantities as opposed to strongly labeled data.

We propose to follow up the DCASE 2018 task 4 [6] and investigate the scenario where large scale detection system exploits the availability of small weakly annotated dataset, a larger unlabeled dataset and an additional training set with strongly annotated synthetic soundscapes. In this task, we focus on SED with time boundaries in domestic environments. For the task, system has to detect the presence of a sound event and predict the onset and offset of the sound event to provide temporal context of when a sound event occurs. We generate strongly annotated synthetic soundscapes using the Scraper tool [12]. Scraper generates a soundscape keeping the description of the foreground and background events and is agnostic to the type of sound events. Furthermore, generation depends upon the sound event specification such as event duration, event time, pitch shift, time stretch etc which allows us to generate multiple versions of audio clips according to our requirements based on a given description. Since generating such strongly labeled synthetic data is feasible on large scale, we provide synthetic data to explore the scientific question - Do we really need real but partially and weakly annotated data or is using synthetic data sufficient? or do we need both for sound event detection? We believe insights learned from this task will be beneficial for the community as such an exploration is novel and will provide a pathway to develop scalable systems for sound event detection.

This manuscript describes the DCASE 2019 task 4 and is organized as follows. Section 2 provides a brief overview of the task description and how the development and evaluation dataset are created. Section 3 describes the baseline system and it's evaluation process for task 4. Section 4 gives an overview of the evaluation submissions. Discussion and Conclusions from the challenge are presented in section 5 and 6

2. TASK DESCRIPTION AND DATASET

2.1. Task description

This task is the follow-up to DCASE 2018 task 4 [6]. Systems are expected to produce strongly labeled output (i.e. detect sound events with a start time, end time, and sound class label), but are provided with weakly labeled data (i.e. sound recordings with only the presence/absence of a sound included in the labels without any

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS Learning to understand audio scenes (ANR-18-CE23-0020) and the French region Grand-Est. Experiments presented in this paper were carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000)

timing information) for training. Multiple events can be present in each audio recording, including overlapping events. As in the previous iteration of this task, the challenge entails exploiting a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set to improve system performance. However, unlike last year, in this iteration of the challenge we also provide an additional training set with strongly annotated synthetic soundscapes. This opens the door to exploring scientific questions around the informativeness of real (but weakly labeled) data versus strongly-labeled synthetic data, whether the two data sources are complementary or not, and how to best leverage these datasets to optimize system performance.

2.2. Development dataset

The dataset for this task is composed of 10 sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. The task focuses on the 10 classes of sound events used in DCASE 2018 task 4 [6]. The dataset for DCASE 2019 task 4 is composed of a subset with real recordings (extracted from Audioset) and a subset with synthetic soundscapes. The subset with real recordings is the same as in DCASE 2018 task 4: the training set remains the same [6] and the validation set is the combination of DCASE 2018 task 4 validation set and evaluation set [10].

2.2.1. *Synthetic soundscapes generation procedure*

The subset with synthetic soundscapes is composed of 10 sec audio clips generated with Scaper [12], a python library for soundscape synthesis and augmentation. Scaper operates by taking a set of foreground sounds and a set of background sounds automatically sequencing them into random soundscapes sampled from a user-specified distribution controlling the number and type of sound events, their duration, signal-to-noise ratio, and several other key characteristics. The foreground events are obtained from the freesound dataset (FSD) [13, 14]. Each sound event clip was verified by a human to ensure that the sound quality and the event-to-background ratio were sufficient to be used as an isolated sound event. We also controlled if the sound event onset and offset were present in the clip. Each selected clip was then segmented when needed to remove silences before and after the sound event and between sound events when the file contained multiple occurrences of the sound event class. The number of unique isolated sound event per class used to generate the subset of synthetic soundscapes is presented in Table 1.

The background textures are obtained from the SINS dataset (activity class ‘other‘) [15]. This particular activity class was selected because it presents a low amount of sound events from the 10 target sound event classes. However, there is no guarantee that these sound event classes are totally absent from the background clips. A total of 2060 unique background clips are used to generate the synthetic dataset.

Scaper scripts are designed such that the distribution of sound events per class, the number of sound events per clip (depending on the class) and the sound event class co-occurrence are similar to that of the validation set composed of real recordings. The synthetic soundscapes are annotated with strong labels automatically generated by Scaper [12].

Class	Unique events	Dev set	
		Clips	Events
Alarm/bell/ringing	190	392	755
Blender	98	436	540
Cat	88	274	547
Dishes	109	444	814
Dog	136	319	516
Electric shaver/toothbrush	56	221	230
Frying	64	130	137
Running water	68	143	157
Speech	128	1272	2132
Vacuum cleaner	74	196	204
Total	1011	2045	6032

Table 1: Class-wise statistics for the synthetic soundscapes development subset

2.3. Evaluation dataset

The evaluation dataset is composed of two subsets. A first subset is composed of audio clips extracted from youtube and vimeo videos under creative common licenses. This subset is used for ranking purposes.

A second subset is composed on synthetic soundscapes generated with Scaper. This subset is used for analysis purposes and its design is motivated by the analysis of last year results [10]. The foreground events are obtained from the FSD [13, 14]. The selection process was the same as described for the development dataset. Background sounds are extracted from youtube videos under creative common license and from the freesound subset of the MUSAN dataset [16]. Audio clips are artificially degraded using Audio Degradation Toolbox [17].

More details about the evaluation set will be provided after the evaluation period.

3. BASELINE

The baseline system has been inspired by Lu, the DCASE 2018 task 4 winner [18]. It uses a mean-teacher model which is a combination of two models: a student model and a teacher model (both have the same architecture). The implementation of Mean teacher model is based on Tarvainen and Valpola work [19]. The student model is the final model, and the teacher model aims to help the student model. The teacher model weights are an exponential moving average of student model weights.

The student model is trained on synthetic and weakly labeled data. The classification cost (binary cross entropy) is computed at frame level on synthetic data and at clip level on weakly labeled data. The teacher model is not trained, its weights are a moving average of the student model (at each epoch). For all data, the teacher model receives the input of the student model with a Gaussian noise added, and helps the student model thanks to a consistency cost (mean-squared error) for strong (frame-level) and weak predictions. During the training, all batches contain unlabeled, weak and strong labeled data. The 4 costs are combined as follow:

$$L(\theta) = L_{class_w}(\theta) + \sigma(\lambda)L_{cons_w}(\theta) + L_{class_s}(\theta_s) + \sigma(\lambda)L_{cons_s}(\theta_s)$$

where θ are the weights of the model, θ_s are the weights of the model without the activation layer.

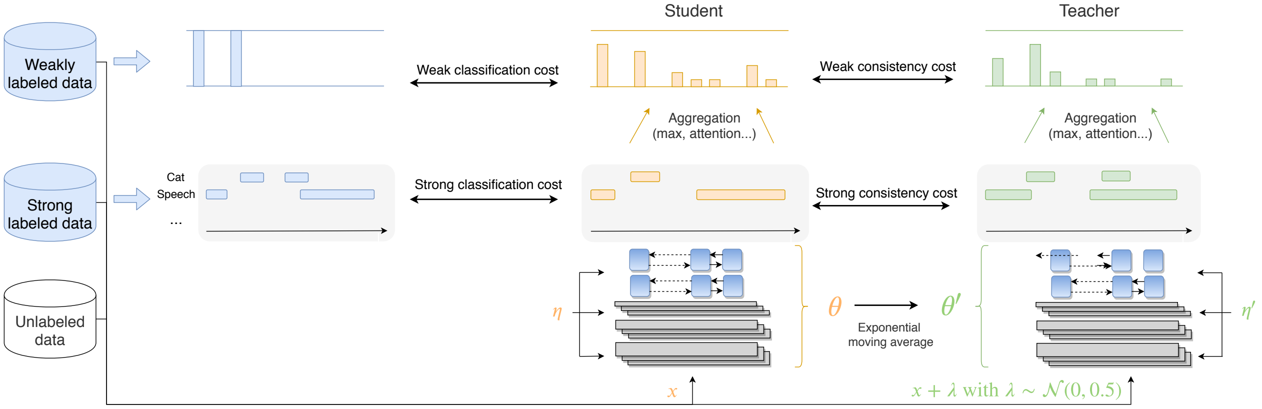

Figure 1: Mean teacher model, θ and θ' are the weights of the student and teacher model respectively. η and η' are noises applied to the different models (here dropout).

The models are a combination of convolutional neural network (CNN) and recurrent neural network (RNN) followed by an aggregation layer (attention here). The output of the RNN gives strong predictions while the output of the aggregated layer gives the weak predictions. The code is open source.[1]

4. SUBMISSION EVALUATIONS

More details about this will be provided after the evaluation period.

5. DISCUSSION

More details about this will be provided after the evaluation period.

6. CONCLUSION

More details about this will be provided after the evaluation period.

7. ACKNOWLEDGMENT

The authors would like to thank the Hamid Eghbal-Zadeh from Johannes Kepler University (Austria) who participated to the initial discussions about this task as well as all participants to the task.

8. REFERENCES

[1] C. Mydlarz, J. Salamon, and J. P. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Applied Acoustics*, vol. 117, pp. 207–218, 2017.

[2] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 373–397.

[3] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

[4] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[5] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," July 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://hal.inria.fr/hal-01850270

[7] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[8] Y. Zigel, D. Litvak, and I. Gannot, "A method for automatic fall detection of elderly people using floor vibrations and soundproof of concept on human mimicking doll falls," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2858–2867, 2009.

[9] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[10] R. Serizel and N. Turpault, "Sound Event Detection from Partially Annotated Data: Trends and Challenges," in *IcETRAN conference*, Srebrno Jezero, Serbia, June 2019. [Online]. Available: https://hal.inria.fr/hal-02114652

[11] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[12] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[1] `https://github.com/turpaultn/DCASE2019_task4/tree/public/baseline`

[13] F. Font, G. Roma, and X. Serra, “Freesound technical demo,” in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 411–412.

[14] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, “Freesound datasets: a platform for the creation of open audio datasets,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[15] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, “The SINS database for detection of daily activities in a home environment using an acoustic sensor network,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 32–36.

[16] D. Snyder, G. Chen, and D. Povey, “MUSAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[17] M. Mauch and S. Ewert, “The audio degradation toolbox and its application to robustness evaluation,” in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 2013, pp. 83–88.

[18] L. JiaKai, “Mean teacher convolution system for dcase 2018 task 4,” DCASE2018 Challenge, Tech. Rep., September 2018.

[19] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, p. 10.